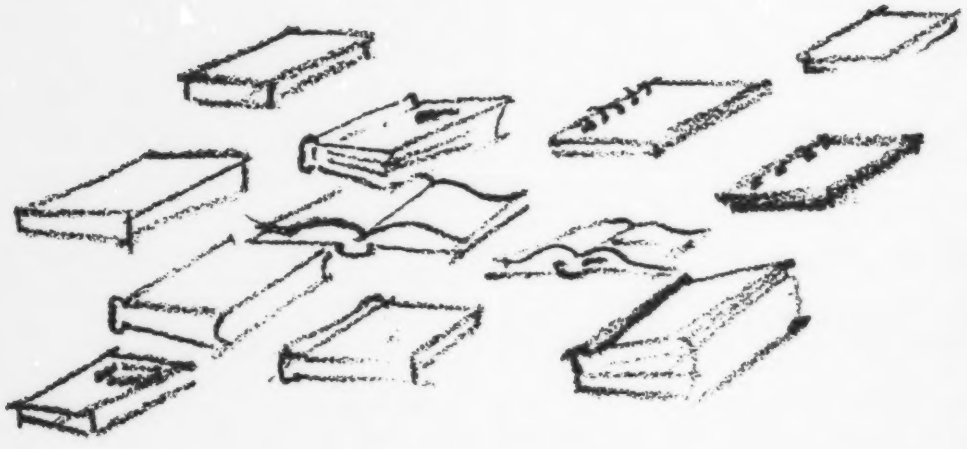
statistiques

DOCUMENTS DE RÉFÉRENCE

RECUEIL DES TEXTES DES PRÉSENTATIONS DU COLLOQUE SUR LES MÉTHODES ET DOMAINES D'APPLICATION DE LA STATISTIQUE 1990



Bureau de
la statistique
du Québec

91-01353/

Pour tout renseignement concernant le BSQ
et les données statistiques qui y sont disponibles,
s'adresser au:

Bureau de la statistique du Québec
117, rue Saint-André
Québec (Québec)
G1K 3Y3
Tél.: (418) 691-2401

ou

Tél.: 1-800-463-4090
(aucuns frais d'appel)

Dépôt légal
Bibliothèque nationale du Québec
Quatrième trimestre 1990
ISBN 2-551-12500-6

© Bureau de la statistique du Québec

Toute reproduction est interdite
sans l'autorisation expresse
du Bureau de la statistique du Québec.

Avant-propos

Ce recueil contient les textes des présentations du colloque "Méthodes et domaines d'application de la statistique 1990" qui s'est tenu à l'Université Laval à Québec en mai 1990, dans le cadre du 58e congrès de l'Association canadienne-française pour l'avancement des sciences. Comme par les années passées, le colloque était parrainé par le Bureau de la statistique du Québec; neuf séances totalisant une cinquantaine de conférenciers ont sollicité la présence et l'enthousiasme de plus de cent vingt participants et participantes de tous les milieux de la statistique.

Nous tenons à remercier tous ceux et celles qui nous ont fait parvenir le texte de leur présentation. Nous espérons que la diffusion de ce recueil permettra aux statisticiens et aux statisticiennes de profiter des expériences et recherches ici démontrées. Le Bureau de la statistique du Québec laisse aux auteurs et auteures la responsabilité entière du contenu de leur texte.

Le directeur général

Luc Bessette

Responsables du colloque 1990 : Louise Guilbault,
statisticienne
Bun Luk Khuong,
directeur
Direction de la méthodologie

Coordonnatrice de cette publication: Françoise Tarte
statisticienne

Pour tout renseignement concernant le contenu de cette publication:
Direction de la méthodologie
Bureau de la statistique du Québec
117, rue St-André
Québec (Québec)
G1K 3Y3
Tél.: (418) 691-2411

Table des matières

OUVERTURE DU COLLOQUE

SAVARD, Guy, Directeur général adjoint, Bureau de la statistique du Québec.
Allocution d'ouverture 3

SESSIONS SUR LA MÉTHODOLOGIE D'ENQUÊTES

ARMSTRONG, John, LEPETIT, Christel, Statistique Canada.
Répartition de l'échantillon pour un plan de sondage à deux phases, avec application aux données fiscales 15

BÉLANGER, Yves, Statistique Canada.
Le remaniement des enquêtes sur les cultures à Statistique Canada 27

BERGERON, Paul, ministère des Transports.
Bilan méthodologique de sondages évaluant la satisfaction des usagers du réseau routier 35

CORRIVEAU, Pierre, ministère du Revenu du Québec.
Projet pilote d'un centre de micrographie 43

DUMAIS, Jean, LAVALLÉE, Pierre, Statistique Canada.
Une approche pour l'échantillonnage et l'estimation d'enquêtes trimestrielles: l'Enquête sur le camionnage pour compte d'autrui 51

LABILLOIS, Tony, Statistique Canada.
Un logiciel d'aide à la sélection des plans d'échantillonnage en contrôle de la qualité 59

LOIGNON, Nathalie, PARÉ, Robert, VEILLETTE, Sylvain, ministère du Revenu du Québec.
Reconstitution des ménages québécois en 1986 à l'aide des déclarations de revenus de 1986 69

LYNCH, Jenny, Statistique Canada.
La méthodologie de la composante aréolaire de l'Enquête nationale sur les fermes 79

SEOANE, Nicole A, ministère de l'Agriculture, des Pêcheries et de l'Alimentation du Québec.
La méthodologie des enquêtes sur le comportement alimentaire de groupes cibles de la population québécoise 89

SESSION SUR LA BIOSTATISTIQUE

ALARY, Michel, BRISSON, Jacques, Université Laval.
Composantes séquentielles des rapports de densité d'incidence 101

BERNARD, Paul Marie, LAPOINTE, Claude. Université Laval.
Calcul des tailles d'échantillons pour la détection d'un rapport de risques, suivant un niveau de confiance et une puissance déterminée 109

BERNIER, Julie, BOISJOLY, Hélène Marie, Centre Hospitalier de l'Université Laval, BERNARD, Paul-Marie, Université Laval.
Utilisation des modèles de régression de Cox dans l'évaluation du taux de survie d'une greffe de cornée 117

ÉMOND, Lise, COURTEAU, Jean-Pierre, Département de santé communautaire de l'Outaouais.
Comment expliquer les disparités entre les taux de mortalité provenant du ministère fédéral Santé et Bien-être social Canada et ceux du ministère provincial de la Santé et des Services sociaux? 125

SESSION SUR LA QUALITÉ ET LA CONFIDENTIALITÉ DES DONNÉES

BOUDREAU, Jean-René, Statistique Canada.
Risque de divulgation (définition et applications) 137

DANEAULT, Sylvie, GUILBAULT, Louise, KHUONG, Bun Luk, Bureau de la statistique du Québec.
Le traitement statistique pour la confidentialité des microdonnées au Bureau de la statistique du Québec 149

LAVALLÉE, Pierre, GRAVEL, Claude, Statistique Canada.
Analyse de la validité et de la fiabilité d'un questionnaire d'opinion 161

PERRON, Sylvain, Statistique Canada.
Le défi posé par les nouvelles technologies 175

SESSIONS SUR L'ANALYSE DE DONNÉES

BÉRARD, Anick, Agriculture Canada.
Analyse statistique des effets de la nature des aliments sur l'activité myoélectrique réticulo-ruminale et sur le comportement chez le veau 189

CARMICHAEL, Jean-Pierre, DAIGLE, Gaétan, RIVEST, Louis-Paul, Université Laval.
Analyse statistique des préférences des canadiens en matière de bière 201

COLIN, Bernard, Université de Sherbrooke, MAC-GIBBON-T, Brenda, Université du Québec à Montréal et Université Concordia, TIBEIRO, Jules, Université de Moncton.
Dépouillement d'une enquête statistique relative aux attitudes des infirmières de la région montréalaise 213

COUSINEAU, Marie-Marthe, CUCUMEL, Guy, Université du Québec à Montréal.
Utilisation de techniques d'analyses factorielles et de classification automatique dans le traitement d'un corpus de données sociologiques 223

CUCUMEL, Guy, Université du Québec à Montréal.
Construction d'une hiérarchie consensus à l'aide d'une ultramétrique centrale 235

LAFOND, Mario, Université Laval.
Modèles linéaires généralisés: analyse des résidus pour des données provenant d'une distribution binomiale 245

LAPOINTE, François-Joseph, Université de Montréal.
La comparaison statistique de classifications hiérarchiques: Évaluation de trois tests de comparaison par permutations 255

MELAGNE, Litchy P., CLÉMENT, Bernard, Dr. REMISE, Claude, École Polytechnique de l'Université de Montréal.
Analyse de courbes de croissance: un exemple en céphalométrie 267

MORISSETTE, Daniel, Université Laval.
Étude des données influentes sous le modèle de régression de Cox 287

PETITJEAN ROGET, Thierry, Institut de recherche en santé et en sécurité du travail, ARSENAULT, André, Institut de cardiologie de Montréal.
Utilisation complémentaire de l'analyse des données (AFC, ACP) et des analyses paramétriques en santé et en sécurité du travail 295

SESSION SUR L'ESTIMATION

BÉLAND, Yves, Statistique Canada.
Étude de la variabilité l'échantillonnage 307

JULIEN, Claude, Statistique Canada.
La possibilité d'estimer le taux d'erreur nette de couverture lors du recensement de la population de 1991 323

LASSARRE, Sylvain, Institut National de recherche dans les transports et leur sécurité (France), ROWENCZYK, Jean-Pierre, Démoscopie (France).
STRATOR, un logiciel généraliste pour le traitement de données d'enquête 335

LEBLOND, Yves, Statistique Canada.
Correction simple pour le biais des estimateurs de variance traditionnels 345

RANGER, Normand, Université de Montréal.
Connaissez-vous S? 353

FORUM POUR UNE ASSOCIATION DE STATISTICIENS

DESBIENS, Christian, ministère des Transports du Québec.
Nouvelle association de statisticiens 365

MONTEGIANI, Mario, ministère des Transports du Québec.
L'Union fait la force 371

TREMBLAY, Victor, Statplus.
Présentation de l'Association internationale des statisticiens d'enquêtes 383

Congrès de l'ACFAS, 1990
5e colloque sur les méthodes et domaines d'application de la statistique

Allocution d'ouverture

par
Guy SAVARD
Directeur général adjoint du BSQ

Le colloque qui s'ouvre maintenant a un caractère un peu spécial. Il marque en effet le cinquième anniversaire de cette activité, et cela mérite d'être souligné de façon particulière.

Il y a six ans, c'était tout un défi que de songer à organiser un premier forum. Au cours des années suivantes ce fut aussi un défi de maintenir et de faire progresser cette activité.

C'est donc d'abord à vous tous qu'il faut souhaiter bon anniversaire, puisque c'est votre colloque et que vous l'avez soutenu par votre participation.

Mais la vitalité d'un colloque scientifique s'appuie sur les efforts de tous ceux qui daignent y présenter les résultats de leurs travaux. Il y a donc une centaine de collègues de la communauté des statisticiens, qu'ils soient présents ou absents aujourd'hui, qu'il faut féliciter d'avoir préparé l'une ou plusieurs des 230 communications présentées à ce colloque annuel depuis ses débuts, y incluant celui qui commence.

Une mention particulière doit aller aux organisateurs, car c'est eux qui ont assuré la mobilisation des énergies et la continuité. L'initiateur et l'organisateur du premier colloque fut Gilbert Émond. Puis ce sont Françoise Tarte, Martine Désilets et Louise Guilbault qui ont pris la relève jusqu'à ce jour. Ces quatre collègues ont droit à nos plus sincères remerciements et à nos félicitations pour tout le travail réalisé.

Un cinquième anniversaire dans l'existence d'une telle activité est une bonne occasion de faire un premier bilan. Les objectifs qu'on s'était fixés au départ ont-ils été atteints? Quels nouveaux objectifs pourraient être envisagés pour l'avenir?

À l'origine, on souhaitait qu'un tel colloque fournisse plusieurs avantages à ses participants:

- l'occasion de faire connaître les travaux en cours au plan des méthodes et des domaines d'application de la statistique;

- la possibilité de se faire connaître et de se faire reconnaître par la communauté de ceux qui oeuvrent dans ce champ d'activité;

- L'OPPORTUNITÉ DE SE CONSULTER ENTRE PAIRS SUR LES PROBLÈMES RENCONTRÉS DANS NOS TRAVAUX;

- CELLE DE SUSCITER DES ÉCHANGES ENTRE LES STATISTICIENS QUI PRATIQUENT LEUR MÉTIER DANS DIFFÉRENTS SECTEURS D'ACTIVITÉ, ENTRE CEUX DES ORGANISATIONS GOUVERNEMENTALES OU PRIVÉES, ET CEUX DU MILIEU UNIVERSITAIRE;

- L'OCCASION DE VALIDER DES MÉTHODES, DE DÉCOUVRIR DE NOUVELLES PISTES ET DE CONNAÎTRE LES DERNIERS DÉVELOPPEMENTS.

QU'EN EST-IL APRÈS CINQ ANS? QUELS RÉSULTATS PEUT-ON OBSERVER, PAR RAPPORT À CES OBJECTIFS?

- ON ENREGISTRE D'ABORD UNE AUGMENTATION CONTINUE DES COMMUNICATIONS D'ANNÉE EN ANNÉE. DE 27 EN 1986, ELLES ONT ATTEINT 74 EN 1989 SOIT UNE AUGMENTATION DE 174 %, APRÈS ÊTRE PASSÉES À 33 EN 1987 ET À 42 EN 1988. AVEC UN NOMBRE RELATIVEMENT ÉLEVÉ DE COMMUNICATIONS, SOIT 53, LA PRÉSENTE ANNÉE SEMBLE CEPENDANT MARQUER UN CERTAIN REPLI.

- Le nombre des communications qui sont publiées par la suite, dans le recueil des textes du colloque, croît aussi d'année en année. Elles étaient au nombre de 16 en 1986, puis sont passées à 19 et 29 en 1987 et 1988 pour atteindre 37 en 1989. La proportion des communications publiées par rapport à celles présentées apparaît cependant instable, fluctuant entre 50 et 69 %.

- Les secteurs et les milieux d'appartenance de ceux qui présentent des communications révèle par ailleurs une variété assez représentative de la réalité. Si on se base sur les communications publiées, on note que le nombre de communications faites par des statisticiens oeuvrant au gouvernement du Québec est relativement stable, aux environs de 7 chaque année, alors que celles en provenance de Statistique Canada sont passées de 6 en 1986 à 13 en 1989. Ce sont les communications des universitaires qui ont connu l'évolution la plus marquée, passant de une en 86 à 16 en 1989. Les communications annoncées pour le présent colloque confirment ces tendances; il y en aura 7 de la fonction publique québécoise, 18 de SC et 25 du milieu universitaire.

- Quant aux sujets des communications, leur variété réflète aussi le large éventail des méthodes et des champs d'application de la statistique. Si l'on compare le colloque de 1986 à celui de l'an dernier, on constate une évolution dans les secteurs dominants de préoccupation. Dans le premier colloque, sur 27 communications, 9 ont porté sur les méthodologies d'enquête, 5 sur l'analyse de données, 3 sur l'exploitation de fichiers administratifs et 2 sur les séries chronologiques. Dans le dernier colloque, sur les 74 exposés, 23 avaient trait à la méthodologie alors que 32 ont porté sur l'analyse des données, laissant loin derrière les séries chronologiques avec 9 présentations et l'exploitation des fichiers administratifs avec 6 communications seulement.

Ces quelques constats généraux permettent sans doute de conclure que les principaux objectifs de votre colloque annuel sont atteints. Votre colloque représente bien le carrefour qu'on a voulu qu'il soit et un carrefour qui s'élargit et se renouvelle d'année en année.

Mais comme dans tout bilan, on peut aussi se demander s'il y a place à des améliorations, si des changements ou des initiatives nouvelles pourraient permettre de faire encore mieux. Vous êtes

ÉVIDEMMENT LES MIEUX PLACÉS POUR EN DÉBATTRE, EN MESURANT L'ÉVOLUTION DE VOS ATTENTES ET EN RÉEXAMINANT LES STANDARDS QUE VOUS VOULEZ VISER. NÉANMOINS, VOUS TROUVEREZ SANS DOUTE UTILE POUR STIMULER VOTRE PROPRE RÉFLEXION, QU'UN OBSERVATEUR EXTÉRIEUR, QUI N'EST PAS DE VOTRE DISCIPLINE ET QUI N'EST PAS NON PLUS UN HABITUÉ DES COLLOQUES SCIENTIFIQUES, SUGGÈRE QUELQUES RÉFLEXIONS EN CE SENS:

- LE POURCENTAGE DES COMMUNICATIONS PUBLIÉES PEUT SUSCITER UNE PREMIÈRE INTERROGATION. ON PEUT EN EFFET S'ÉTONNER EN CONSTATANT QUE POUR TROIS ANNÉES SUR QUATRE DE 40 À 50 % DES COMMUNICATIONS PRÉSENTÉES AU COLLOQUE NE SE SONT PAS RENDUES JUSQU'À LEUR PUBLICATION DANS LE RECUEIL DES TEXTES. S'AGIT-IL D'UN MANQUE DE PRÉPARATION, D'UN MANQUE DE TEMPS OU D'UN MANQUE DE SOUCI DU TRAVAIL FINI DE LA PART D'UN BON NOMBRE DE PRÉSENTATEURS, OU ENCORE D'UN CHANGEMENT DE PRIORITÉ OU D'UNE DIMINUTION DE L'INTÉRÊT CHEZ CES DERNIERS PLUS LE TEMPS PASSE APRÈS LE COLLOQUE? IL Y A SANS DOUTE UN PEU DE TOUT CELA. MAIS N'Y A-T-IL PAS LIEU DE CROIRE QUE PLUSIEURS DES OBJECTIFS DU COLLOQUE, RAPPELÉS PLUS TÔT, SONT MAL SERVIS PAR UN TAUX AUSSI ÉLEVÉ DE NON PUBLICATION? TOUTE AMÉLIORATION À CET ÉGARD DÉPEND ÉVIDEMMENT DE L'AUTO-DISCIPLINE DES COMMUNICATEURS EUX-MÊMES. CELLE-CI SERAIT CEPENDANT FAVORISÉE S'IL

ÉTAIT DORÉNAVANT CONSIDÉRÉ COMME LA NORME QUE TOUTE COMMUNICATION COMPORTE L'ENGAGEMENT DE PUBLICATION ET QUE LE TEXTE DES COMMUNICATIONS DOIT ÊTRE TRANSMIS DANS LES DEUX MOIS SUIVANT LE COLLOQUE. LES ACTES DU COLLOQUE POURRAIENT AINSI ÊTRE DIFFUSÉS DÈS L'AUTOMNE ET LEUR CONTENU GARDER UN INTÉRÊT PLUS MARQUÉ.

- UN DES OBJECTIFS DU COLLOQUE ÉTAIT DE FAVORISER LES ÉCHANGES ENTRE LES STATISTICIENS DES DIFFÉRENTS MILIEUX, NOTAMMENT ENTRE CEUX DES DIVERS SECTEURS DE L'ADMINISTRATION PUBLIQUE, ENTRE CEUX DES DIFFÉRENTES UNIVERSITÉS ET PLUS PARTICULIÈREMENT ENCORE ENTRE LES STATISTICIENS DE CES DEUX MILIEUX, SANS OUBLIER CEUX DU SECTEUR PRIVÉ. L'ÉVOLUTION DE LA PARTICIPATION AUX COMMUNICATIONS MONTRE BIEN UNE PRÉSENCE ET UNE IMPLICATION DE PLUS EN PLUS REPRÉSENTATIVES DE L'ÉVENTAIL DES SECTEURS ET MILIEUX D'APPARTENANCE. PEUT-ÊTRE QU'UN PROGRÈS SIMILAIRE DANS L'IMPLICATION POURRAIT MAINTENANT SE FAIRE AU PLAN DE L'ORGANISATION DU COLLOQUE. DEPUIS SES DÉBUTS, LE RÔLE ESSENTIEL DU BSQ, EN TANT QUE PARRAIN DU COLLOQUE, A ÉTÉ DE FAIRE ASSUMER PAR DES MEMBRES DE SON PERSONNEL L'ORGANISATION ANNUELLE ET LA CONTINUITÉ DE CELUI-CI. LA COLLABORATION DE CERTAINS COLLÈGUES DANS LEURS MILIEUX RESPECTIFS LEUR A ÉTÉ, BIEN ENTENDU, INDISPENSABLE. MAIS

N'Y A-T-IL PAS LIEU MAINTENANT DE RENDRE PLUS COLLÉGIALE OU COMMUNAUTAIRE LA RESPONSABILITÉ DE PLANIFIER, D'ORGANISER ET DE FAIRE LE SUIVI DU COLLOQUE ANNUEL? N'Y AURAIT-IL PAS LÀ UNE MARQUE ENCORE PLUS TANGIBLE DE L'ENGAGEMENT DES DIFFÉRENTS MILIEUX À L'ÉGARD DU COLLOQUE, UNE GARANTIE ACCRUE DE SA CONTINUITÉ ET UN STIMULANT ADDITIONNEL AU DÉVELOPPEMENT DES ÉCHANGES? LE PARRAINAGE DU BSQ S'ACCOMODERAIT FORT BIEN D'UN TEL ÉLARGISSEMENT DE LA PARTICIPATION DU MILIEU.

- LA RECONNAISSANCE DE L'EXCELLENCE PAR SES PAIRS CRÉE GÉNÉRALEMENT UNE ÉMULATION QUI CONTRIBUE À LA RECHERCHE DE CETTE EXCELLENCE PAR LES INDIVIDUS, TOUT EN FAVORISANT L'IDENTIFICATION DE CEUX-CI À LA COMMUNAUTÉ SCIENTIFIQUE D'APPARTENANCE. IL N'Y A PAS DE DOUTE AUSSI QUE L'ÉVÉNEMENT QUI EST L'OCCASION DE CETTE RECONNAISSANCE DE L'EXCELLENCE SE TROUVE AINSI À PRENDRE UNE IMPORTANCE TOUTE PARTICULIÈRE POUR SES DIVERSES CLIENTÈLES. NE SERAIT-IL PAS INTÉRESSANT QUE VOTRE COLLOQUE ANNUEL SOIT L'OCCASION DE LA REMISE DE QUELQUES PRIX D'EXCELLENCE AUX COLLÈGUES MÉRITANTS? PAR EXEMPLE: PRIX D'EXCELLENCE POUR L'INNOVATION MÉTHODOLOGIQUE, OU ENCORE PRIX D'EXCELLENCE POUR L'IMPACT SOCIAL D'UNE CONTRIBUTION. UNE TELLE INITIATIVE CONTRIBUERAIT SANS DOUTE À CONSOLIDER L'IMAGE DE MARQUE DE VOTRE COLLOQUE. NUL DOUTE AUSSI QUE PLUSIEURS DES

ORGANISATIONS OU REGROUPEMENTS D'ORGANISATIONS DONT VOUS FAITES PARTIE ACCEPTERAIENT D'APPUYER DE MANIÈRE TANGIBLE LA REMISE DE TELS PRIX.

VOICI DONC TROIS SUGGESTIONS D'AMÉLIORATIONS OU D'INITIATIVES NOUVELLES QUI PERMETTRAIENT DE FAIRE ENCORE MIEUX, DE FAIRE ENCORE PLUS. TOUTE ORGANISATION A BESOIN POUR MAINTENIR SA VITALITÉ DE SE FIXER DE NOUVEAUX OBJECTIFS QUAND CEUX QU'ELLE POURSUIVAIT ONT ÉTÉ ATTEINTS OU SONT EN VOIE DE L'ÊTRE. LES SUGGESTIONS QUE J'AI FORMULÉES ONT POUR BUT DE VOUS INCITER À VOUS FIXER VOUS-MÊMES DE NOUVEAUX DÉFIS POUR LES 5 PROCHAINES ANNÉES, DES DÉFIS RÉALISTES DONT VOUS POURREZ CÉLÉBRER AVEC FIERTÉ LA RÉALISATION LORS DE VOTRE DIXIÈME ANNIVERSAIRE.

D'ICI LÀ, JE VOUS SOUHAITE BON COLLOQUE, TANT POUR CELUI QUI COMMENCE QUE POUR CEUX DES PROCHAINES ANNÉES.

14 MAI 1990

Répartition de l'échantillon pour un plan de sondage à deux phases, avec application aux données fiscales

J. Armstrong et C. Le Petit[1]

Dans le cadre du projet de remaniement des enquêtes-entreprises à Statistique Canada, des modifications à l'enquête annuelle produisant des estimations à partir des données fiscales ont été introduites. La nouvelle enquête se sert d'un plan de sondage à deux phases. Pour la première phase, les strates sont définies en utilisant entre autres la classification type des industries, avec deux chiffres de précision (CTI2). Les strates de la deuxième phase utilisent la classification type des industries, avec quatre chiffres de précision (CTI4). On veut publier des estimations au niveau CTI4 avec un niveau de précision donné, représenté par un certain coefficient de variation au niveau CTI4. Cet article examine le problème de la répartition de l'échantillon de cette enquête.

1. Introduction

Dans le cadre du projet de remaniement des enquêtes-entreprises qui se déroule à Statistique Canada depuis 1985, des modifications à l'enquête utilisant les données fiscales de Revenu Canada ont été introduites. L'objectif de l'enquête est de produire des estimations pour des variables financières pour les domaines définis par le croisement de la Classification Type des Industries avec quatre chiffres de précision (CTI4) et des provinces. Seulement deux chiffres de la CTI sont codés avec certitude à Revenu Canada. Avant le début du projet de remaniement des enquêtes-entreprises, Statistique Canada effectuait une enquête utilisant les données fiscales de Revenu Canada avec un plan de sondage à une phase. Les strates étaient définies utilisant les provinces et la taille (en terme de revenu brut). Le nombre d'unités qui tombaient dans une cellule CTI4 *x* province était aléatoire et dépendait de la distribution des CTI4. Par conséquent, la précision des estimations variait d'une industrie à l'autre.

Afin de rendre la précision des estimations au niveau CTI4 *x* province plus uniforme, la nouvelle enquête utilise un plan de sondage à deux phases. On sélectionne l'échantillon de première phase constitué de déclarants fiscaux à Revenu Canada en utilisant des strates définies à l'aide des variables CTI2, province et taille.

Un déclarant fiscal peut posséder plusieurs entreprises. Pour chaque entreprise d'un déclarant fiscal dans l'échantillon de première phase, on détermine la CTI4 à Statistique Canada. Lors de la sélection de l'échantillon de deuxième phase, on utilise des strates définies au niveau CTI4 *x* province *x* taille. La sélection de l'échantillon à chaque phase se fait par échantillonnage de Bernouilli (aussi appelé échantillonnage de Poisson). Les tailles d'échantillon réalisées en se servant de cette méthode sont aléatoires. Choudhry, Lavallée et Hidiroglou (1989) donne plus de détails sur le plan de sondage.

Le but de cet article est d'examiner le problème de la répartition de l'échantillon. Une contrainte de précision pour une variable d'enquête, le revenu brut, a été spécifiée pour chaque domaine CTI4 *x* province. Par contre, il n'y a aucune contrainte de précision au niveau des strates de taille. La méthode de répartition optimale (répartition de Neyman) convient à cette situation. On note que d'autres méthodes de répartition, par exemple la méthode de répartition exponentielle décrite par Bankier (1988), seraient préférables s'il y avait des contraintes de précision associées aux strates de taille.

La solution au problème de répartition pour un plan de sondage à deux phases lorsqu'il y a une seule contrainte globale sur l'estimation est bien connue et est donnée par Cochran (1977). Récemment, Smith (1989) a étudié le problème de répartition quand il y a une contrainte de précision sur une combinaison linéaire d'estimations de domaines. Il est à noter que le problème que l'on considère dans le présent article est différent parce qu'il y a une contrainte de précision pour chacun des domaines.

---

1 John Armstrong, méthodologiste principal, Division des méthodes d'enquêtes-entreprise, Christel Le Petit, méthodologiste, Division des méthodes d'enquêtes-sociales, Statistique Canada, Ottawa, K1A 0T6

Dans cet article, on donne premièrement une formulation du problème de la répartition conjointe des échantillons de première et de deuxième phase. La solution de cette formulation, que l'on appelle la méthode exacte, donne la répartition optimale. On donne en deuxième lieu une méthode d'approximation de la répartition optimale. On présente par la suite les résultats de quelques comparaisons entre la méthode d'approximation et la méthode exacte. On termine par une brève conclusion.

2. Méthode exacte

Dans cette section, on décrit le problème de la répartition conjointe de première et de deuxième phase. La solution à ce problème donne la répartition optimale pour le plan de sondage à deux phases. Afin d'expliquer le problème, on considère une cellule CTI2 x province contenant N unités. Chaque unité tombe dans l'une de H cellules CTI4 x province. La cellule CTI2 x province contient G strates de première phase CTI2 x province x taille et $G \times H$ strates de deuxième phase CTI4 x province x taille. Appelons N_g le nombre d'unités dans la strate CTI2 x province x taille g et υ_g la fraction de sondage de première phase correspondante. De même, appelons N_{gh} le nombre d'unités et υ_{gh} la fraction de sondage de deuxième phase, respectivement, pour la strate CTI4 x province x taille gh. On dénote le total, dans la population, de la variable étudiée pour la cellule CTI4 x province h par Y_h et l'estimation de Horwitz-Thompson du total par $\hat{Y}_h$.

Choudhry, Lavallée et Hidiroglou (1989) montrent qu'en conditionnant sur la taille d'échantillon obtenue, l'échantillonnage de Bernouilli est équivalent à l'échantillonnage aléatoire simple dans l'hypothèse ou la probabilité de sélection de chaque unité dans une strate est la même. Ainsi la variance de $\hat{Y}_h$ pour le plan de sondage à deux phases peut être représentée par

$$V_h = \sum_g \left(\frac{1}{\upsilon_g \cdot \upsilon_{gh}} - 1 \right) \cdot A_{gh} + \sum_g \left(\frac{1}{\upsilon_g} - 1 \right) \cdot B_{gh},$$

où

$$A_{gh} = N_{gh} \cdot S^2_{gh},$$

$$B_{gh} = \left(\frac{N_g - N_{gh}}{N_g - 1} \right) \cdot \left(\frac{Y^2_{gh}}{N_{gh}} - S^2_{gh} \right),$$

et S^2_{gh} est la variance de la population dans la strate de deuxième phase CTI4 x province x taille gh.

Le coût de la sélection d'une unité dans l'échantillon de première phase est K_1, peu importe la strate dans laquelle se trouve l'unité, tandis que le coût de la sélection d'une unité dans l'échantillon de deuxième phase est K_2, peu importe la strate. Le problème de la répartition conjointe des échantillons de première et de deuxième phase revient à trouver le minimum de la fonction de coût

$$F = K_1 \cdot \sum_g \upsilon_g \cdot N_g + K_2 \cdot \sum_g \sum_h \upsilon_g \cdot \upsilon_{gh} \cdot N_{gh}, \qquad (1)$$

par rapport à υ_g, $g = 1, 2, \ldots, G$, et υ_{gh}, $g = 1, 2, \ldots, G$, $h = 1, 2, \ldots, H$ sous les contraintes

$$\sum_g \left(\frac{1}{\upsilon_g \cdot \upsilon_{gh}} - 1 \right) \cdot A_{gh} - \sum_g \left(\frac{1}{\upsilon_g} - 1 \right) \cdot B_{gh} \le C_h^2 \cdot Y_h^2, \quad h = 1, 2, \ldots, H, \qquad (2)$$

$$0 < \upsilon_g \le 1, \quad g = 1, 2, \ldots, G,$$

$$0 < \upsilon_{gh} \le 1, \quad g = 1, 2, \ldots, G, \quad h = 1, 2, \ldots, H,$$

où C_h dénote le coefficient de variation cible pour le domaine d'intérêt h, équivalent à une cellule CTI4 x province.

Puisque certaines des contraintes de (2) vérifieront l'inégalité à la solution, des méthodes d'optimisation non linéaire doivent être utilisées pour résoudre ce problème. Toutefois, comme Wu (1989) l'a suggéré, le problème peut être sensiblement simplifié en le séparant en deux étapes principales résolues itérativement. D'abord, on trouve le minimum de (1) par rapport à υ_g, $g = 1, 2, \ldots, G$ sous réserve d'un ensemble de valeurs pour les fractions de sondage de deuxième phase. Cette étape exige l'emploi de techniques d'optimisation non linéaires. La deuxième étape consiste à trouver le minimum de (1) par rapport aux fractions de sondage de deuxième phase, conditionnellement aux valeurs des fractions de sondage de première phase obtenues à la première étape. L'obtention du minimum peut se faire indépendamment pour chaque $h = 1, 2, \ldots, H$. On recommence alors la première étape et ainsi de suite jusqu'à convergence.

Grâce à ce processus itératif, on obtient la répartition en solutionnant une série de problèmes d'optimisation non linéaire, dont chacun ne comporte que G variables. Dénotons par $\upsilon_g^{(i)}$ et $\upsilon_{gh}^{(i)}$ les estimations des valeurs optimales de υ_g et de υ_{gh} obtenues après i itérations. Au début de l'itération $i+1$, les changements de variables définis par $X_g^{(i+1)} = 1/(\upsilon_g^{(i+1)} - 1)$ peuvent être utilisés pour transformer le problème d'optimisation en un problème comportant une fonction objective non linéaire et des contraintes linéaires convenant à l'application de méthodes de programmation non linéaire. En guise de valeurs de départ pour la première itération on pose $\upsilon_{gh}^{(0)} = 1$, $g = 1, 2, \ldots, G$, $h = 1, 2, \ldots, H$.

Plus précisément, chaque itération comporte les étapes suivantes:

(i) Recherche du minimum de

$$F = \sum_g \left(N_g + \frac{K_2}{K_1} \sum_h \upsilon_{gh}^{(i-1)} \cdot N_{gh} \right) / (X_g^{(i)} + 1)$$

par rapport à $X_g^{(i)}$, $g = 1, 2, \ldots, G$, sous les contraintes

$$C_h^2 \cdot Y_h^2 - \sum_g \left(\frac{X_g^{(i)} + 1}{\upsilon_{gh}^{(i-1)}} - 1 \right) \cdot A_{gh} - \sum_g X_g^{(i)} \cdot B_{gh} \geq 0, \quad h = 1, 2, \ldots H$$

$$X_g^{(i)} \geq 0, \quad g = 1, 2, \ldots, G.$$

(ii) Calcul de $\upsilon_g^{(i)} = 1/(X_g^{(i)} + 1)$, $g = 1, 2, \ldots G$

(iii) Recherche du minimum, pour $h = 1, 2, \ldots H$, de

$$F_h = \sum_g \upsilon_g^{(i)} \cdot \upsilon_{gh}^{(i)} \cdot N_{gh}$$

par rapport à $\upsilon_{gh}^{(i)}$, $g = 1, 2, \ldots G$ sous les contraintes

$$C_h^2 \cdot Y_h^2 - \sum_g \left(\frac{1}{\upsilon_g^{(i)} \cdot \upsilon_{gh}^{(i)}} - 1 \right) \cdot A_{gh} - \sum_g \left(\frac{1}{\upsilon_g^{(i)}} - 1 \right) \cdot B_{gh} \geq 0,$$

$$0 < \upsilon_{gh}^{(i)} \leq 1, \quad g = 1, 2, \ldots, G.$$

On verra dans la prochaine section que la solution de l'étape (iii) est analytique. De plus il est à noter que lors de la première étape de chaque itération, les contraintes sont convexes et la fonction objective est concave. Donc, il s'agit d'un problème de programmation convexe, une forme particulière de problème d'optimisation non linéaire. Pour ce genre de problème, on peut montrer qu'il n'y a qu'une solution locale. Par conséquent, pour trouver la solution globale du problème, il suffit de trouver une solution locale.

3. Méthode d'approximation

Dans cette section, on décrit une méthode de répartition qui donne une approximation de la répartition optimale. En supposant que toutes les fractions de sondage de deuxième phase sont égales à un, on calcule premièrement une approximation de la répartition optimale de l'échantillon de première phase. Ensuite, on fait la répartition de l'échantillon de deuxième phase, étant donné les fractions de sondage de première phase. Le coût du prélèvement d'une unité dans l'échantillon à chaque phase d'échantillonnage étant constant pour l'ensemble des strates, la réduction du coût équivaut à la réduction de la taille de l'échantillon à chaque étape de cette approche.

La première étape de la méthode consiste en la résolution approximative du problème de répartition pour un plan de sondage à une phase. Cette étape revient à trouver le minimum, indépendamment pour chaque h, de

$$F^{(h)} = \sum_g \upsilon_{g|h} \cdot N_g \tag{3}$$

par rapport à $\upsilon_{g|h}$, $g = 1, 2, \ldots, G$, sous les contraintes

$$\sum_g \left(\frac{1}{\upsilon_{g|h}} - 1 \right) \cdot (A_{gh} + B_{gh}) \leq C_h^2 \cdot Y_h^2, \tag{4}$$

$$0 < \upsilon_{g|h} \leq 1, \quad g = 1, 2, \ldots, G. \tag{5}$$

On peut montrer que le minimum de (3) est obtenu quand les membres de (4) sont égaux, de sorte que le problème défini par (3), (4) et (5) équivaut à trouver le minimum du lagrangien

$$L^{(h)} = \sum_g \upsilon_{g|h} N_g + \lambda \cdot \left[C_h^2 \cdot Y_h^2 - \sum_g \left(\frac{1}{\upsilon_{g|h}} - 1 \right) \cdot (A_{gh} + B_{gh}) \right].$$

La solution de ce problème est donnée par:

$$\upsilon_{g|h} = ((A_{gh} + B_{gh}) / N_g)^{1/2} \cdot \sum_g ((A_{gh} + B_{gh}) \cdot N_g)^{1/2} / \left(C_h^2 \cdot Y_h^2 + \sum_g (A_{gh} + B_{gh}) \right). \tag{6}$$

Si une ou plusieurs des fractions de sondage données par (6) sont plus grandes que un, on peut les poser égales à un et résoudre un problème modifié comportant un nombre réduit de strates. Cette méthode correspond au procédé de sur-répartition mentionné dans Cochran (1977). Enfin, on pose υ_g égal à la valeur la plus élevée de l'ensemble $\{\upsilon_{g|h}, \; h = 1, 2, \ldots H\}$ pour $g = 1, 2, \ldots, G$, ce qui permet de satisfaire les contraintes pour chaque h.

Étant donné les fractions de sondage de première phase, on peut facilement trouver les fractions de sondage optimales de deuxième phase. Appelons n_{gh} le nombre d'unités choisies à la première phase se trouvant dans la strate de deuxième phase CTI4 x province x taille gh. Notons que n_{gh} ne sera pas nécessairement égal à son espérance, qui est donnée par $\upsilon_g \cdot N_{gh}$. Le problème de la répartition de l'échantillon de deuxième phase peut être considéré comme le problème consistant à minimiser

$$F_h = \sum_g \upsilon_{gh} \cdot n_{gh} \tag{7}$$

par rapport à υ_{gh}, $g = 1, \ldots, G$, sous les contraintes

$$\sum_g \frac{1}{\upsilon_g} \cdot \left(\frac{1}{\upsilon_{gh}} - 1 \right) \cdot A_{gh} \leq M_h. \tag{8}$$

$$0 < \upsilon_{gh} \leq 1, \quad g = 1, \ldots, G, \tag{9}$$

où

$$M_h = C_h^2 \cdot Y_h^2 - \sum_g \left(\frac{1}{\upsilon_g} - 1 \right) \cdot (A_{gh} + B_{gh}).$$

Il est facile de montrer que (7) atteint un minimum quand l'égalité de la contrainte (8) est vérifiée. Par conséquent le problème équivaut à trouver le minimum du lagrangien

$$L_h = \sum_g \upsilon_{gh} \cdot n_{gh} + \lambda \cdot \left(M_h - \sum_g \frac{1}{\upsilon_g} \cdot \left(\frac{1}{\upsilon_{gh}} - 1 \right) \cdot A_{gh} \right)$$

par rapport à υ_{gh}, $g = 1, \ldots, G$. et λ, sous les contraintes

$$0 < \upsilon_{gh} \leq 1, \quad g = 1, \ldots, G.$$

Ce problème peut être résolu de façon analytique. La solution est:

$$\upsilon_{gh} = [A_{gh}/(\upsilon_g \cdot n_{gh})]^{1/2} \cdot \sum_g (n_{gh} \cdot A_{gh}/\upsilon_g)^{1/2} / D_h \tag{10}$$

où

$$D_h = C_h^2 \cdot Y_h^2 + \sum_g \left(\frac{1}{\upsilon_g} \right) \cdot A_{gh} - \sum_g \left(\frac{1}{\upsilon_g} - 1 \right) \cdot (A_{gh} + B_{gh}).$$

Notons que le problème de répartition demeure sans solution à moins que D_h soit positif.

Si certaines valeurs de υ_{gh} sont plus grandes que un, on peut évidemment utiliser le procédé de sur-répartition décrit plus haut. En pratique, N_{gh}, Y_h et S_{gh}^2 sont inconnus, et il est nécessaire d'utiliser des estimations fondées sur les données de l'année la plus récente disponible. Il est à noter que, pour un h donné, si on remplace n_{gh} dans (10) par son espérance par rapport à la première phase d'échantillonnage, $\upsilon_g \cdot N_{gh}$, on obtient la solution analytique de l'étape (iii) de la méthode exacte.

4. Comparaison des deux méthodes

On effectue ici une comparaison entre la méthode exacte et la méthode approximative décrites dans les sections 2 et 3. On considère en autres le cas particulier où G=2 et H=2. Une analyse des situations où elles sont équivalentes et des situations où la méthode approximative donne de moins bons résultats que la méthode exacte est effectuée. Une étude des coûts engendrés par l'utilisation de la méthode approximative est aussi présentée.

Comme dans la section 2, on effectue un changement de variables afin de transformer un problème d'optimisation d'une fonction linéaire sur une région définie par des fonctions non linéaires en un problème d'optimisation non-linéaire sur une région définie par des fonctions linéaires.

De plus, afin de simplifier la discussion, on étudie le cas où G=2 et H=2. Posons:

$$X_1 = \frac{1}{\upsilon_1} - 1 \qquad X_2 = \frac{1}{\upsilon_2} - 1.$$

Comme H=2 on a deux problèmes à résoudre pour la solution approximative.

Pour h=1 on doit premièrement minimiser

$$F^{(1)} = \frac{N_1}{X_{111}+1} + \frac{N_2}{X_{211}+1} \tag{11}$$

sous les contraintes

$$X_{111}.(A_{11}+B_{11}) + X_{211}.(A_{21}+B_{21}) \leq C_1^2.Y_1^2 \tag{12}$$

$$X_{111} \geq 0 \quad \text{et} \quad X_{211} \geq 0 \tag{13}$$

Deuxièmement, pour h=2, on doit minimiser

$$F^{(2)} = \frac{N_1}{X_{112}+1} + \frac{N_2}{X_{212}+1}$$

sous les contraintes

$$X_{112}.(A_{12}+B_{12}) + X_{212}.(A_{22}+B_{22}) \leq C_2^2.Y_2^2, \tag{14}$$

$$X_{112} \geq 0 \quad \text{et} \quad X_{212} \geq 0 \tag{15}$$

Une fois que X_{111}, X_{112}, X_{211} et X_{212} qui minimisent $F^{(1)}$ et $F^{(2)}$ sont obtenus, on calcule υ_1 et υ_2 de la façon suivante

$$X_1 = \min(X_{111}, X_{112}), \qquad X_2 = \min(X_{211}, X_{212}),$$

$$\upsilon_1 = \frac{1}{X_1+1} \quad \text{et} \quad \upsilon_2 = \frac{1}{X_2+1}.$$

Pour la solution exacte, la fonction à minimiser dans le cas G=2 et H=2 est donnée par

$$F = \left(1 + \frac{K_1}{K_2}\right).\left\{\frac{N_1}{X_1+1} + \frac{N_2}{X_2+1}\right\} \tag{16}.$$

Celle-ci est équivalente à $F^{(1)}$ et à $F^{(2)}$ à une constante près. On minimise F simultanément sous les contraintes de $F^{(1)}$ et de $F^{(2)}$ soit:

$$X_1.(A_{11}+B_{11}) + X_2.(A_{21}+B_{21}) \leq C_1^2.Y_1^2$$

$$X_1.(A_{12}+B_{12}) + X_2.(A_{22}+B_{22}) \leq C_2^2.Y_2^2$$

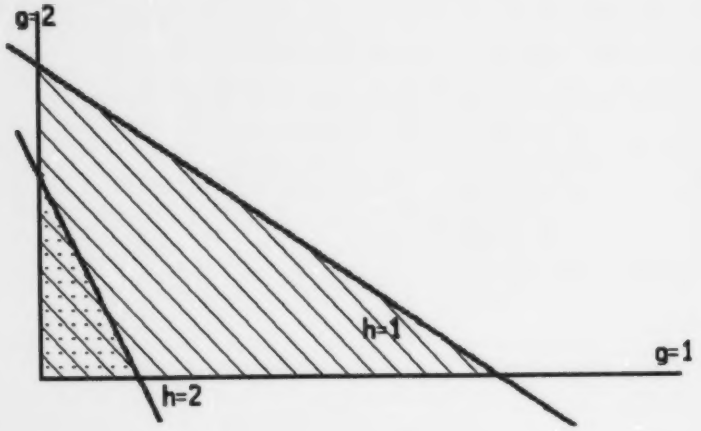
$$X_1 \geq 0 \quad \text{et} \quad X_2 \geq 0$$

La région définie par les contraintes étant convexe il s'agit d'un problème de programmation convexe. Dans cette situation, si on trouve un minimum local il s'agira aussi d'un minimum global. Une fois que X_1 et X_2 qui minimisent F sont trouvés, on obtient υ_1 et υ_2 de la façon suivante:

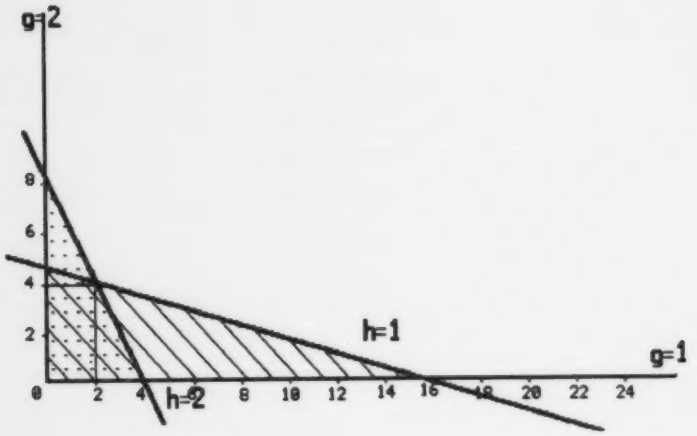
$$\upsilon_1 = \frac{1}{X_1+1} \quad \text{et} \quad \upsilon_2 = \frac{1}{X_2+1}$$

On voit donc que la fonction à minimiser pour les deux approches est la même, la différence étant dans le nombre de contraintes. Ainsi, pour l'approche approximative, on trouve deux solutions sous deux séries de contraintes puis on prend le minimum des deux solutions, alors que pour la solution exacte on minimise la même fonction sous toutes les contraintes simultanément.

On peut représenter les contraintes graphiquement. Les schémas 1 et 2 représentent deux cas avec deux ensembles de contraintes différents.



Type de région
Schéma 1

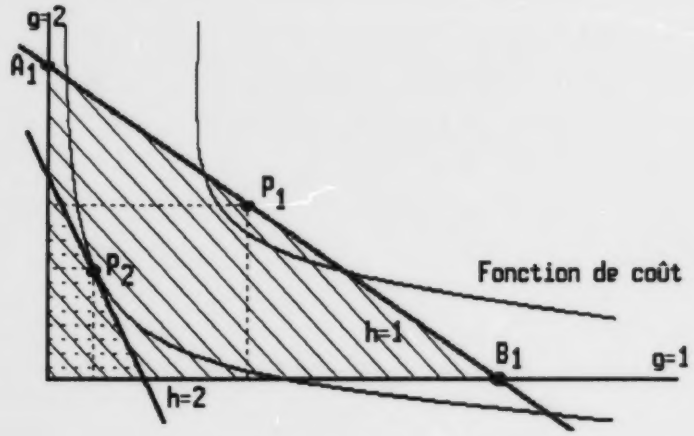


Type de région
Schéma 2

Sur ces schémas les strates de taille sont représentées par les axes. Les droites h=1 et h=2 représentent les contraintes liées à la précision pour chaque h. Les contraintes pour le premier problème de la solution approximative (équations (12) et (13)) définissent la région hachurée. Celles du deuxième problème (équations (14) et (15)) définissent la région identifiée par des points. Quant à la solution exacte, elle consiste à minimiser F sous toutes les contraintes, soit sur le domaine représenté par l'intersection des régions hachurées et à points.

Les deux schémas représentent deux types de régions différents. Dans le cas du schéma 1 la contrainte liée à h=2 est plus restrictive que celle liée à h=1 pour toutes les valeurs de X_1 et X_2, c'est à dire qu'elle est plus proche des axes. Dans le cas du schéma 2, aucune des deux contraintes n'est plus restrictive que l'autre pour toutes les valeurs de X_1 et X_2. La contrainte liée à h=1 est plus restrictive pour les valeurs de X_1 allant de 0 à 2 et pour les valeurs de X_2 allant de 4 à 4.5, alors que la contrainte liée à h=2 est plus restrictive pour les valeurs de X_1 allant de 2 à 4 et les valeurs de X_2 allant de 0 à 4.

Étudions en premier lieu la solution approximative dans le cas où on a une région similaire à celle du schéma 1, soit la situation où une contrainte est plus restrictive que l'autre pour toutes les valeurs de X. L'équation (11) définit une famille de fonction pour chaque h et chaque membre de cette famille correspond à une valeur fixe de F_h. On veut trouver la fonction de cette famille qui est aussi loin que possible de l'origine tout en restant dans la région pour maximiser X_1 et X_2. Quelques-unes des fonctions de cette famille sont représentées au schéma 3.

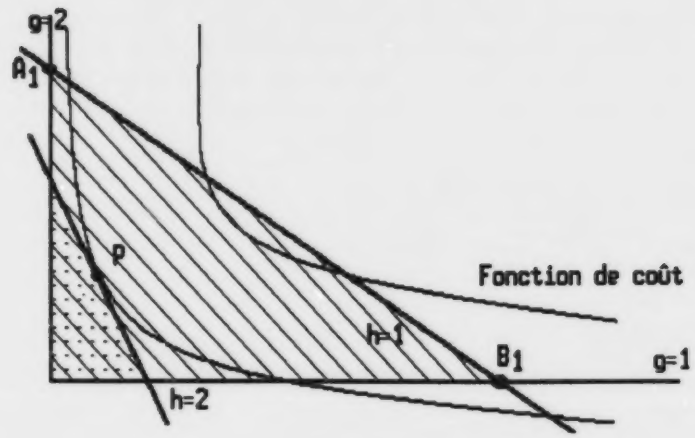


Méthode approximative

Schéma 3

La solution du premier problème (h=1) de la solution approximative est représentée par P_1 alors que la solution du deuxième problème (h=2) est représentée par P_2. Une fois ces deux points obtenus on prend la fraction de sondage maximum qui satisfait les deux contraintes, soient les X_1 et X_2 minimum. Dans notre cas, le point représentant ce minimum est P_2 car il a une plus petite abcisse et une plus petite ordonnée.

Si on utilise maintenant l'approche exacte sur ce même type de domaine, la première itération consiste à minimiser la même fonction F sur le domaine défini par l'intersection des zones hachurées et à points (voir schéma 4).

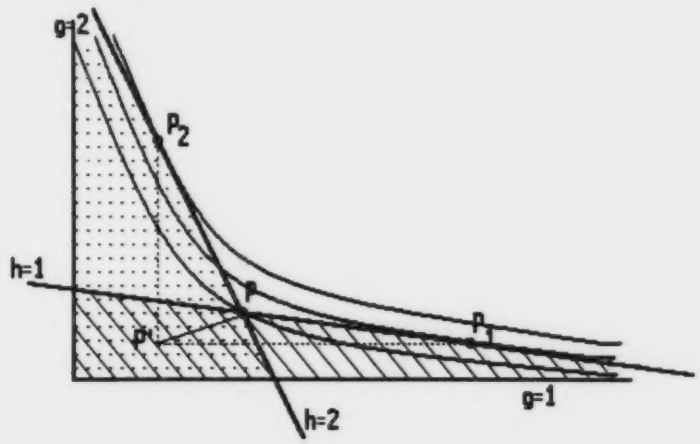


Méthode exacte. Première itération
Schéma 4

Le point optimal est représenté par le point P, qui correspond exactement au point P_2 du schéma 3. Donc, dans le cas où le même h détermine la fraction de sondage maximale: $\{\upsilon_{g|h}, h = 1, \ldots, H\}$ pour tous les g, la solution approximative et la première itération de la solution exacte donnent la même solution. Une fois les υ_g obtenus à la première itération, on calcule les υ_{gh} selon la méthode décrite à la section 3.

Pour la deuxième itération, la fonction à minimiser ainsi que les contraintes changent car les valeurs de υ_{gh} sont différentes de 1. On obtient alors d'autres fractions de sondage de première phase qui sont utilisées pour calculer un autre ensemble de fractions de deuxième phase, et ainsi de suite jusqu'à convergence. Notons qu'à chaque itération la valeur de la fonction de coût F diminue.

Il y a cependant des régions où ce n'est pas le même h qui détermine la fraction de sondage maximale pour chaque g. On se trouve alors dans la situation présentée par le schéma 5.



Comparaison des méthodes exacte et approximative
Schéma 5

Le point P_1 représente la solution du premier problème de la méthode approximative et le point P_2 la solution du deuxième problème. Quand on prend le minimum de ces deux points on obtient le point P' qui est la solution de la méthode approximative. La méthode exacte quant à elle nous donne le point P comme solution, qui est plus loin de l'origine que P' et correspond donc à des fractions de sondage inférieures. Dans cette situation la méthode exacte donne de meilleurs résultats dès la première itération. La distance entre P et P' montre qu'il y a une perte d'efficacité due à l'utilisation de la méthode approximative. On calcule ensuite les υ_{gh} pour la deuxième phase, et on continue le processus itératif jusqu'à convergence.

La discussion des paragraphes précédents peut être généralisée aux dimensions g=1,...,4 et h=1,...,H pour chaque combinaison CTI2 *x* province afin de correspondre à la situation de l'enquête. Afin de comparer les deux approches, on discrimine les cellules CTI2 *x* province en trois groupes selon le critère suivant:

GROUPE A: Le même h détermine la fraction de sondage maximale pour les quatre g.

GROUPE B: Le même h détermine la fraction de sondage maximale pour trois g parmi quatre.

GROUPE C: Le même h ne détermine pas la fraction de sondage maximale pour plus de deux g parmi quatre.

La taille de la population est de 922 638 unités pour l'ensemble du Canada. L'échantillon qui a été réparti selon la méthode approximative comprend 97 716 unités. La répartition des unités de l'échantillon entre les groupes est de 17% des unités dans le groupe A, 28% dans le groupe B et 55% dans le groupe C. Pour le Québec, elle est de 18% dans le groupe A, 35% dans le groupe B et 47% dans le groupe C.

Afin d'étudier l'ampleur du gain obtenu en utilisant la solution exacte au lieu de la solution approximative, l'étude suivante a été faite pour le Québec. Pour chaque CTI2 non-vide, les fractions de sondage de première et deuxième phase ont été calculées avec les deux approches. Pour obtenir les coûts associés à ces fractions de sondage il faut déterminer le rapport entre K_1 et K_2 les coûts de sélection de première et deuxième phase. En fait les coûts de sélection sont identiques à la première et à la deuxième phase, soit environ sept dollars. Cependant l'échantillon de première phase étant longitudinal, une unité sélectionnée une année donnée reste dans l'échantillon de première phase les années subséquentes, alors que l'échantillon de deuxième phase est sélectionné chaque année indépendamment des années précédentes. Le rapport entre les coûts de première phase et de deuxième phase a été fixé pour l'étude à $K_2 / K_1 = 5$, ce qui suppose qu'une unité de première phase reste en moyenne cinq ans dans l'échantillon avant de mourir. Lors de la comparaison des deux méthodes avec les données du Québec, on a posé $K_1 = 7/5$ et $K_2 = 7$. On note que les coûts d'entretien de l'échantillon de première phase (révision cyclique des codes CTI4, éliminations des unités mortes, etc...) n'ont pas été pris en compte dans notre étude.

Afin de solutionner le problème de programmation non linéaire relié à la méthode exacte, on a utilisé des sous-routines de la librairie IMSL (1987) qui sont basées sur l'algorithme de Schittkowski (1985). Cet algorithme est sensible aux points de départ. La solution approximative a été utilisée comme point de départ, et des essais avec d'autres points de départ ont également été effectués. L'algorithme n'a pas réussi à atteindre la solution lors de tous les essais.

Le tableau 1 présente, pour chaque groupe, la différence nette entre les deux coûts et la différence en terme de pourcentage de perte obtenue utilisant la solution approximative par rapport à la solution exacte.

Groupe	Coût solution exacte	Coût solution approximative	Différence nette	Différence %
A	18 800 67	21 460.32	2 659.66	14.9%
B	32 606.80	34 670.16	2 063.36	6.3%
C	57 984.02	59 811.21	1 827.18	3.1%
TOTAL	109 391.49	115 941.69	6 550.20	6.0%

Coût des deux méthodes pour les groupes A, B et C
Tableau 1

Le gain total obtenu grâce à l'utilisation de la solution exacte est de 6% seulement. Le gain maximal avec la solution exacte a été réalisé pour le groupe A (15%), soit le groupe où les deux méthodes donnaient le même résultat après une itération. Ceci implique que le gain provient des itérations subséquentes. Pour les deux autres groupes, bien que la méthode exacte donne de meilleurs résultats après une itération, les itérations suivantes n'ont permis qu'une amélioration inférieure à celle du groupe A. Ceci peut être expliqué par la nature des données. Il y a moins de groupes de taille vides pour les domaines du groupe A. Par contre pour les strates des groupes B et C, on avait davantage de domaines vides pour des groupes de taille. Or, ces groupes de taille vides peuvent être considérés comme des contraintes supplémentaires, qui limitent ainsi les gains pouvant être obtenus avec la méthode exacte.

Même si le coût total avec la méthode exacte est toujours inférieur à celui de la méthode approximative, l'examen des résultats a montré que le coût pour la première phase était souvent supérieur avec la méthode exacte et que celui de la deuxième phase était inférieur. Ceci vient du fait que le rapport entre K_2 et K_1 a été fixé à 5. Il revient donc à moins cher de prendre davantage d'unités à la première phase, ce qui donne une meilleure base de sondage pour la deuxième phase et permet de sélectionner moins d'unités.

La simulation a été effectuée pour le Québec, soit la deuxième plus grande province en terme de population d'entreprises au Canada et le gain total en terme de coût grâce à l'utilisation de la méthode exacte est seulement de 6%. Comme on l'a vu, la solution exacte semble donner de meilleurs résultats pour les domaines complets et homogènes. Ceci laisse à penser que si on appliquait cette méthode aux petites provinces, les gains en terme de coût ne seraient pas importants.

5. Conclusion

Cette étude nous a permis de comparer deux approches qui peuvent être utilisées pour résoudre le problème de la répartition de l'échantillon des données fiscales. L'une de ces approches, la solution approximative, est facile à implanter et a déjà été utilisée pour la répartition de l'échantillon de 1990. L'autre approche, la solution exacte, offre une réduction en terme du coût, mais pose également des problèmes d'application avec des données réelles. On constate que l'algorithme arrête souvent avant de trouver une solution et semble très sensible aux points de départ qui lui sont offerts. De plus, lorsque les domaines ne sont pas complets pour chaque groupe de taille, ce qui est souvent le cas avec les données réelles, le gain obtenu peut être considéré comme marginal. Ces dernières remarques nous incitent donc à continuer d'utiliser la solution approximative pour la répartition de l'échantillon des données fiscales.

BIBLIOGRAPHIE

BANKIER, M.D. (1988). Power allocations: determining sample sizes for subnational areas. *The American Statistician*, 42, 174-177.

CHOUDHRY, G.H., LAVALLÉE, P., HIDIROGLOU, M. (1989). Two-phase sample design for tax data. Communication présenté lors de la réunion annuelle de l'American Statistical Association.

COCHRAN, W.G. (1977). Sampling Techniques, (3^{e} éd.). New York : John Wiley.

SCHITTKOWSKI, K. (1985). NLPQL: A Fortran subroutine solving constrained nonlinear programming problems. Annals of Operation Research 5, 485-500.

IMSL(1987). Math/Library FORTRAN Subroutines for Mathematical Applications. Houston: IMSL Inc.

SMITH, P.J. (1989). Is two-phase sampling really better for estimating age composition?. Journal of the American Statistical Association, 84, 916-921.

WU, J. (1989). Procès-verbal de la réunion du Comité consultatif des méthodes statistiques, le 3-4 avril, 1989. Document interne, Statistique Canada.

LE REMANIEMENT DES ENQUÊTES SUR LES CULTURES À STATISTIQUE CANADA

Communication présentée au colloque

Méthodes et domaines d'application de la statistique 1990

dans le cadre du 58ième congrès de l'ACFAS

Yves Bélanger

Division des méthodes d'enquêtes-entreprises
Statistique Canada
Ottawa (Ontario) K1A 0T6

Introduction

Chaque année, Statistique Canada publie une grande variété de renseignements statistiques portant sur les cultures au pays. La majorité de ces renseignements proviennent d'un système composé de sept enquêtes. Celles-ci existent depuis environ 1910 et le coeur en est un échantillon volontaire, non-probabiliste, d'environ 10,000 fermes appelé le Panel. En 1989, on a pris la décision de remanier le plan de sondage du Panel afin de lui donner une base probabiliste, plus défendable au point de vue statistique.

Le nouveau plan de sondage devait satisfaire les trois exigences fondamentales suivantes. Premièrement, il devait être simple, en raison des ressources méthodologiques limitées, et afin qu'il soit facilement compréhensible par les analystes. Deuxièmement, il devait également être robuste, puisqu'il sera utilisé au cours des trois prochaines années, et flexible, puisque le Panel comprend trois types d'enquêtes différents. Finalement, le tout devait être facile à traiter, étant donné que chaque enquête doit être réalisée dans un court intervalle de temps, soit environ un mois.

Pour répondre à ces exigences, on a eu recours à une base de sondage de type liste. De cette base, on a sélectionné un échantillon aléatoire stratifié d'une taille moyenne d'environ 10,000 fermes pour chaque enquête. Les sections qui suivent décriront en détail les enquêtes touchées ainsi que le nouveau plan de sondage.

1. Description des enquêtes

Les enquêtes qui composent le Panel sont au nombre de sept et se répartissent en trois types distincts: il y a une enquête sur les intentions d'ensemencement des superficies cultivables, trois enquêtes sur les stocks de grain sur les fermes et trois enquêtes sur le rendement des cultures. Le tableau 1 (page suivante) montre comment se répartissent ces enquêtes durant l'année, selon le mois de référence de chaque enquête.

Comme on peut le voir, l'enquête sur les intentions d'ensemencement a lieu au début du printemps, au mois de mars. Le cycle des enquêtes sur les stocks de grain commence en décembre, alors que les stocks sont à leur maximum sur les fermes, continue au mois de mars de l'année suivante et se termine en juillet avec les stocks à leur niveau le plus bas. Finalement, la première enquête sur les rendements, en août, permet d'obtenir des données préliminaires concernant le rendement des cultures; l'enquête de septembre fournit la plupart des données finales pour l'est du pays, tandis que l'enquête de novembre fournit l'ensemble des données finales pour l'ouest et le reste des données finales pour l'est.

Il est à noter que bien que ces trois dernières enquêtes portent sur les rendements, on s'intéresse surtout à la production, qui est égale au produit du rendement et de la superficie ensemencée. Les chiffres sur les superficies sont obtenus de l'Enquête nationale sur les fermes, qui est l'enquête agricole la plus importante de Statistique Canada. Cette enquête ne faisant pas partie du groupe

d'enquêtes touchées par le remaniement, on l'a mise entre parenthèses dans le tableau 1.

Tableau 1
Calendrier des enquêtes sur les cultures

Janvier	
Février	
Mars	- Intentions d'ensemencement
	- Stocks
Avril	
Mai	
Juin	- (Enquête nationale sur les fermes)
Juillet	- Stocks
Août	- Rendement
Septembre	- Rendement
Octobre	
Novembre	- Rendement
Décembre	- Stocks

La collecte des données, qui se faisait autrefois par la poste, s'effectue maintenant par entrevues téléphoniques à partir des bureaux régionaux de Statistique Canada et s'étend sur une période de cinq jours. Les questionnaires sont acheminés au bureau-chef d'Ottawa où les données sont saisies, traitées et analysées, et les résultats sont publiés dans le mois qui suit. Ces procédures n'ont pas été modifiées suite au remaniement du plan de sondage.

2. Construction de la base de sondage

Comme on l'a mentionné dans l'introduction, la base de sondage est de type liste. Elle couvre toutes les provinces du Canada à l'exception de Terre-Neuve, dont l'activité agricole insuffisante ne justifie pas son inclusion dans ce groupe d'enquêtes. On peut distinguer trois étapes principales dans la construction de la base:

a) construction d'une base préliminaire:

La première étape a consisté en la construction d'une base préliminaire utilisant les données du recensement de l'agriculture de 1986 et de l'Enquête nationale sur les fermes (ENF) de 1988 et de 1989. On a commencé par prendre toutes les fermes du recensement, au nombre de 292,438; on a ensuite mis à jour les données de 33,854 d'entre elles qui ont été échantillonnées dans l'ENF, soit en 1988 ou en 1989. Les données recueillies sur les superficies ensemencées et la taille des cheptels ont alors remplacé les données correspondantes du recensement. Finalement, 1,021 fermes ont été ajoutées à la base. Ces fermes ont été soit manquées au recensement, soit achetées par des nouveaux venus dans le domaine agricole, et ont été détectées grâce à la base aréolaire de l'ENF. L'addition de ces fermes porte donc à 293,459 le nombre total de fermes dans la base préliminaire.

b) exclusions:

La deuxième étape a consisté d'abord à exclure de la base préliminaire les catégories de fermes suivantes: les fermes institutionnelles, les fermes situées sur des réserves indiennes, les pâturages communautaires et les fermes appartenant à des corporations à opérations multiples. Diverses considérations pratiques ont présidé à ce choix. Dans un deuxième temps, on a exclu une partie des fermes qui ont cessé d'être exploitées depuis le recensement, soit toutes celles que l'on a été en mesure d'identifier. Ensuite, afin de réduire davantage la taille de la base de sondage, on a exclu les petites fermes selon la définition suivante. Dans l'est du pays, c'est à dire en Ontario, au Québec et dans les Maritimes, une petite ferme a été définie comme toute ferme dont la superficie cultivable est inférieure à 60 acres. Dans l'ouest, c'est à dire en Colombie-Britannique et dans les provinces des Prairies, où les fermes sont en général plus grosses, le critère a été de 120 acres.

Il est à noter que dans l'est on a inclus la superficie en foin dans la définition de superficie cultivable. En effet, à cause de la rotation des cultures, un champ de foin, une année donnée, peut être semé d'une culture comme l'avoine ou l'orge l'année suivante. Dans le but d'éviter la détérioration de la base de sondage au fil du temps, on a jugé qu'il était préférable de tenir compte de ce fait. Dans l'ouest, par contre, les pratiques agricoles sont différentes et ce genre de cas est rare.

Après cette étape, le nombre de fermes dans la base de sondage a été réduit à 179,232.

c) couverture:

Dans cette troisième et dernière étape, on a calculé la couverture de la base, après exclusions, par rapport à la base préliminaire. On a considéré trois variables: la superficie cultivable, telle que définie auparavant, la taille du cheptel bovin et la taille du cheptel porcin. Les chiffres ont montré que la couverture de la superficie cultivable au niveau provincial pouvait être aussi basse que 79% (en Nouvelle-Ecosse). De même, pour la taille du cheptel porcin, la couverture n'était que de 17% en Colombie-Britannique, entre autres. On a jugé qu'il était souhaitable d'avoir une couverture minimale de 95% pour chacune des trois variables, dans chacune des provinces. Il y a deux raisons principales pour lesquelles nous voulons une couverture élevée de la taille des cheptels bovins et porcins: premièrement, la présence de ces animaux est reliée à la présence de stocks de grains sur les fermes; deuxièmement, les grosses exploitations de bétail ou de porcs ont le potentiel économique pour acheter ou louer d'importantes superficies de terre cultivable. En incluant ces fermes dans la base de sondage, on se protège donc contre la détérioration de celle-ci au fil du temps.

Pour obtenir la couverture désirée, on a tout simplement ramené dans la base de sondage les fermes contribuant le plus aux trois variables d'intérêt, et qu'on avait précédemment exclues parce qu'elles étaient trop petites. Ceci a été fait indépendamment pour chacune des trois variables.

Après cette étape, la base de sondage est terminée et compte maintenant 207,313 fermes, soit un peu plus de 70% de la taille de la base préliminaire.

3. Stratification

La stratification est effectuée indépendamment à l'intérieur de chacune des provinces, sauf pour les Maritimes qui sont considérées comme une seule province. Dans chacune des provinces, donc, on retrouve deux niveaux de stratification. Le premier niveau est géographique et correspond à chacun des districts agricoles. Un district agricole est habituellement défini comme un ensemble de divisions de recensement contiguës ayant des pratiques agricoles similaires. Au Québec, ce sont plutôt des regroupements de régions agricoles. On compte jusqu'à 20 districts dans une province (en Saskatchewan). Chacune des provinces maritimes est considérée comme un district pour les besoins de la stratification et de la répartition de l'échantillon.

Le deuxième niveau de stratification est constitué de strates de taille à l'intérieur de chacun des districts. On compte trois strates: en général, on peut dire que la strate 1 représente les fermes les plus importantes en terme de superficie cultivable, la strate 2 représente les fermes moyennes et la strate 3 les petites fermes. A ce stade-ci on a utilisé la même définition de superficie cultivable dans l'est et dans l'ouest du pays, soit celle qui exclut la superficie de foin. Ceci a eu comme effet de mettre les fermes de foin de l'est dans la strate 3. Cette strate contiendra donc des petites fermes ayant un certain potentiel de développement au niveau des cultures, soit les fermes avec de grandes étendues de foin et celles où l'on retrouve des cheptels importants, malgré leur faible superficie cultivable.

Lorsqu'on a voulu déterminer les bornes des strates en procédant à une stratification de type hiérarchique avec un progiciel couramment utilisé, on a obtenu des strates très débalancées où la strate 3, par exemple, comprenait plus de 75% des fermes. Ceci a été causé par le fait qu'il y a beaucoup de variabilité dans la taille des grosses fermes. Habituellement, on règle le problème en regroupant les plus grosses fermes dans une strate auto-représentative, ainsi appelée parce que chacune de ces fermes se verra alors sélectionner avec certitude dans l'échantillon. On n'a pas retenu cette solution car la majorité de ces fermes se retrouvent déjà dans les strates auto-représentatives de l'ENF et on a jugé que leur fardeau de réponse serait alors trop élevé.

La solution retenue a été d'attribuer un pourcentage fixe des fermes à la strate 3, et de délimiter ensuite les strates 1 et 2 en utilisant la méthode du cumulatif de la racine carrée. En pratique on a attribué 40% des fermes à la strate 3, et les strates 1 et 2 se sont retrouvées avec respectivement 5% et 55% des fermes environ.

4. Répartition et sélection de l'échantillon

Pour répartir l'échantillon dans chacune des strates, on a fait face à deux contraintes principales. Premièrement, pour des raisons pratiques on voulait conserver les tailles d'échantillon du Panel, qui varient selon la province, bien sûr, mais aussi selon l'enquête; deuxièmement, on désirait que seules les strates 1 et 2 soient échantillonnées pour les trois enquêtes sur les rendements. Dans ce dernier cas, ceci devrait permettre, pour une taille d'échantillon donnée, de concentrer cet échantillon sur des fermes plus grosses et par conséquent de produire des estimations plus précises pour des cultures "rares" comme les cultures spécialisées dans l'ouest, par exemple. L'hypothèse sous-jacente, qui a déjà été vérifiée, est que le rendement ne dépend pas vraiment de la taille de la ferme. Le tableau 2 illustre les tailles d'échantillon visées pour le Québec.

Tableau 2
Tailles d'échantillon initiales pour le Québec

Intentions	Stocks			Rendements*		
	déc.	mars	juil.	août	sept.	nov.
1680	1860	1550	1550	1190	1920	1920

*strates 1 et 2 seulement.

Pour satisfaire la première contrainte on a choisi d'adopter le concept d'échantillon-maître, i.e. un échantillon principal duquel on sélectionne les sous-échantillons correspondant à chacune des enquêtes. Pour faire suite à notre deuxième contrainte, on a décidé d'allouer un pourcentage fixe de l'échantillon dans la strate 3, celle des petites fermes, soit 20%. Etant donné, comme on l'a vu précédemment, que 40% du total des fermes se retrouvent dans cette strate, ceci nous assure que le poids moyen d'une petite ferme sera seulement le double du poids moyen global dans la province. Ceci répond à une autre préoccupation que nous avions, soit de tenter d'éviter que les poids des petites fermes soient trop élevés.

Suite à ces décisions, on calcule facilement la taille de l'échantillon-maître au niveau d'une province: elle est égale à 1.25 multiplié par la taille maximale d'une enquête de rendement, soit 1.25 * 1920 = 2400 pour le Québec dans l'exemple précédent.

On doit maintenant répartir chaque total provincial à travers les strates. On le fait de la manière suivante: premièrement, on divise l'échantillon total entre l'ensemble des strates 1 et 2 combinées, d'une part, et l'ensemble des strates 3, d'autre part, dans la proportion 80%-20%; deuxièmement, on répartit les nombres obtenus dans chacun des districts selon la méthode optimale de Neyman, en utilisant la variable "superficie cultivable". Il ne reste ensuite qu'à déterminer la taille d'échantillon de la strate 1 et celle de la strate 2, séparément dans chacun des districts, en utilisant encore une fois la méthode de Neyman.

L'échantillon-maître est ensuite sélectionné de manière aléatoire simple à l'intérieur de chacune des strates. Ceci constitue l'avant-dernière étape. La dernière étape doit consister en la sélection des sous-échantillons correspondant à chacune des enquêtes. Pour ce faire, on désirait développer une méthode facile à effectuer avec soutien méthodologique minimal, et qui permette un contrôle simple du chevauchement entre sous-échantillons. La solution retenue consiste à diviser l'échantillon-maître en dix "répliques" de tailles égales. Ces répliques sont formées de manière aléatoire de telle sorte qu'une ferme appartienne à une seule réplique et que l'ensemble des dix répliques couvre l'échantillon-maître au complet. Au niveau de la province, ainsi qu'au niveau de la strate, les tailles des répliques diffèrent d'au plus une unité.

Pour sélectionner un sous-échantillon de 60% de l'échantillon-maître, par exemple, il suffit donc de sélectionner n'importe quel groupe de six répliques. Les tailles d'échantillon originales ont été ajustées, soit à la hausse, soit à la baisse, de manière à représenter un nombre entier de répliques. En guise d'illustration, le tableau 3 montre les nouvelles tailles d'échantillon pour le Québec, où l'on a indiqué entre parenthèses le nombre de répliques sélectionnées.

Tableau 3
Tailles d'échantillon finales pour le Québec

Intentions	Stocks			Rendements*		
	déc.	mars	juil.	août	sept.	nov.
1680	1680	1440	1680	1152	1920	1920
(7)	(7)	(6)	(7)	(6)	(10)	(10)

*strates 1 et 2 seulement.

Conclusion

Le nouveau plan de sondage a été conçu et mis en oeuvre entre les mois d'octobre 89 et février 90. La première enquête à utiliser ce nouveau plan de sondage a été l'enquête sur les intentions d'ensemencement, qui a eu lieu en mars 90, et cette première tentative s'est révélée une réussite complète. Pour l'enquête suivante, celle portant sur les stocks de mars, on a sélectionné un échantillon-test de deux provinces avec le nouveau plan de sondage, parallèlement à l'échantillon régulier provenant de l'ancien Panel. On prévoit utiliser uniquement le nouveau plan de sondage pour l'enquête sur les stocks de décembre 90, qui marque le début du cycle des enquêtes sur les stocks.

En ce qui a trait aux enquêtes sur les rendements, nos intentions sont d'utiliser le nouveau plan de sondage pour la première enquête, celle du mois d'août, ainsi que pour les enquêtes subséquentes. Il est donc pratiquement assuré que la nouveau plan de sondage sera complètement opérationnel dès 1991.

BILAN MÉTHODOLOGIQUE DE SONDAGES ÉVALUANT LA SATISFACTION DES USAGERS DU RÉSEAU ROUTIER

Paul Bergeron, statisticien
Ministère des Transports

Septembre 1990

RÉSUMÉ

Deux sondages postaux évaluant la satisfaction des usagers du réseau routier à l'entretien du ministère des Transports du Québec ont été réalisés par la firme SEGMA de Lavalin en collaboration avec la Société (anciennement Régie) de l'assurance automobile du Québec pour le compte du ministère des Transports du Québec. La présentation portera sur les facteurs déterminants ayant permis des taux de réponse supérieurs à 60% ainsi que du suivi méthodologique de la part du Ministère. Quelques résultats de ces sondages seront aussi présentés.

1. INTRODUCTION

Le ministère des Transports du Québec a un mandat principal d'assurer la sécurité et le confort des usagers du réseau routier. Pour s'acquitter de cette tâche importante, le Ministère procède, selon un cycle de gestion standard de planification, programmation, réalisation et évaluation. Les sondages dont il est question s'insèrent justement dans la dernière phase du cycle précité. L'évaluation du niveau de satisfaction des usagers du réseau routier représente un facteur déterminant quant aux orientations à prendre étant donné l'importance de la clientèle pour le Ministère.

À cette fin, deux sondages postaux ont été réalisés par la firme SEGMA de Lavalin (SEGMA est maintenant intégrée à la firme CROP) en collaboration avec la Société (anciennement la Régie) d'assurance automobile du Québec (SAAQ) pour le compte du Ministère. Puisque des résultats étaient désirés pour chacune des 15 directions régionales du Ministère, ces sondages ont été stratifiés par celles-ci. La Direction de la programmation via son Service d'évaluation des programmes a dirigé le projet. Le rôle du Service de la statistique et de l'économie des transports de la Direction de la recherche consistait à assurer l'expertise méthodologique du projet.

Les sujets couverts par ces sondages sont nombreux et ont été choisis en fonction de leur lien direct avec les utilisateurs du réseau routier. Les éléments suivants ont fait l'objet de questions: la sécurité, l'état du réseau, la signalisation, l'entretien d'hiver, le service de renseignement, les travaux, la congestion, le plan de transport de la région de Montréal, les abords de route, les haltes routières, la publicité le long des routes, les priorités et le budget du Ministère.

Cette communication traitera principalement du bilan méthodologique de ces sondages qui furent couronnés de succès. Des taux de réponse souvent supérieurs à 60% dans les différentes régions du Ministère méritent d'être signalés. Le tableau 1 présente ces taux de réponse. (Voir tableau 1 - Annexe A) Même si en soi un taux de réponse élevé ne garantit pas la qualité des résultats, il n'en demeure pas moins important. Ce texte se divise en quatre parties: la problématique, les facteurs déterminants, quelques résultats du premier sondage et une brève conclusion.

2. PROBLÉMATIQUE

Lors de l'élaboration de ce projet de nombreuses questions d'ordre méthodologique furent soulevées. Le choix de confier la réalisation des sondages à une firme privée a conditionné l'importance de l'effort à fournir pour assurer l'atteinte des objectifs du Ministère. Cette affirmation se justifie par le fait que la firme en question est externe au Ministère et qu'à ce titre elle doit être sensibilisée à la problématique interne. Par contre, l'avantage d'un tel choix est justement le "regard neuf" posé par celle-ci. Beaucoup d'énergies ont donc été consacrées à la conception du questionnaire, puisqu'il représente l'intermédiaire qui assure le contact avec le répondant et que nous voulions au besoin répéter la démarche avec ce même questionnaire.

Un des premiers problèmes rencontrés fut de définir ce qu'est "l'usager de la route" et ensuite de fixer un choix sur une base de sondage. En pratique, il est souhaitable de considérer les deux à la fois. Le fichier des détenteurs de permis de conduire de la SAAQ a exercé un attrait considérable pour servir de base de sondage.

La rédaction d'un devis descriptif du travail à effectuer par la firme représentait un travail clé dans le projet. Nous avions donc résolu de rédiger un devis très détaillé garantissant les meilleures chances de satisfaction de notre part. La contribution de Madame Françoise Tarte du Bureau de la statistique du Québec en la matière nous a été utile, car nous nous sommes inspirés largement de documents produits par elle au B.S.Q. et d'un exemple de devis déjà utilisé dans un contexte similaire dans un autre ministère.

Puisque l'état du réseau routier fluctue présumément avec les saisons, il a été décidé de réaliser deux sondages. Les sondages ont été réalisés au printemps et à l'automne 1989 avec des échantillons différents, la répétition visait à connaître les variations dans la perception des usagers du réseau. Comme on le sait, l'état du réseau est au pire au printemps après les ravages de l'hiver et au mieux à l'automne après un été de travaux routiers. Il faut d'ailleurs faire la distinction entre l'état du réseau routier et la perception ou la satisfaction qu'en ont les usagers. La période idéale de référence en aurait été une où l'état du réseau routier fut moyen. Selon nous, en pratique, une telle période n'existe pas. Nous avons exclu réaliser les sondages en été ou en hiver pour différentes raisons. En été, parce que beaucoup de gens s'absentent et que les travaux routiers sont en cours, en hiver, parce que de récentes intempéries auraient pu biaiser les résultats.

Lors de ce projet, les quelques contraintes rencontrées ont mis en exergue la réussite méthodologique de ces sondages. La nécessité de respecter, les dispositions de la loi 101 qui nous obligent à utiliser un questionnaire en langue française lors de l'envoi initial. Il ne s'agit pas ici de décrier l'existence d'une telle loi, mais de souligner l'impact négatif de cette contrainte sur les taux de réponse parmi les détenteurs de permis de conduire allophones (principalement les anglophones). Ceux-ci avaient la possibilité d'obtenir sur demande un questionnaire en langue anglaise. Le fait que les répondants n'étaient pas tenu de répondre au questionnaire selon une loi de la

statistique, comme celles dont le B.S.Q. et Statistique Canada bénéficient, a aussi pu avoir une influence à la baisse sur le nombre de réponses obtenues. Toutefois, cette influence a assurément été mineure comparativement à la précédente.

Le nombre d'intervenants tant à l'interne qu'à l'externe concernés par le projet fut aussi une contrainte, l'impact a surtout été d'allonger les délais car l'organisation logistique devait se faire en respectant les disponibilités et possibilités de chacun.

3. FACTEURS DÉTERMINANTS

Plusieurs éléments ont concouru à l'obtention de taux de réponse élevés pour une enquête postale de cette envergure. Surtout lorsque l'on considère qu'une seule "carte-rappel" a été utilisée pour améliorer la réponse. Certains éléments étaient liés au sujet étudié ou à la base de sondage (le fichier des détenteurs de permis de la SAAQ) et donc moins sous notre contrôle, mais qu'il importe tout de même de mentionner. D'autres éléments comme le moment de la collecte, le devis descriptif ou le questionnaire, représentent des facteurs déterminants qui ont été contrôlés adéquatement. Ce projet a donné lieu à une collaboration fructueuse où chacun s'est bien acquitté de son rôle, "méthodologiquement" parlant.

Nous croyons que plusieurs des nombreux détenteurs de permis sélectionnés au hasard (3540 à chaque sondage) se sont montrés intéressés par la démarche et se sont sentis concernés personnellement par le sujet étudié. Beaucoup de gens ont fait des commentaires à cette fin. Le rôle fondamental du gouvernement dispensateur de services y prend tout son sens.

Puisque le réseau routier à l'étude est celui à l'entretien du ministère des Transports, il a fallu porter une attention particulière à la définition de celui-ci. Cette distinction nous a semblé très importante car les délimitation de juridiction sont rarement indiquées aux usagers. Voici, tiré du questionnaire utilisé, la définition précitée:

"Autoroutes: toutes les autoroutes portent un numéro inférieur à 100 ou supérieur à 399;

Routes numérotées: toutes les routes numérotées ont un numéro variant entre 100 et 399;

Routes rurales: les routes rurales comprennent toutes les routes "non-numérotées" à l'exception des rues et ruelles des villes et villages".

Malgré cette définition élaborée, il nous semble qu'elle a été bien comprise car les résultats sont cohérents par catégorie de route à ce que l'on s'attendait.

Le fichier des détenteurs de permis utilisé comme base de sondage, grâce à la collaboration de la SAAQ, fut très avantageux. Les liens étroits entre les mandats du Ministère et de la SAAQ renforcent la pertinence d'une collaboration

entre nos organisations respectives. Ces liens sont d'autant plus forts que nous nous intéressions, en partie, aux questions de sécurité routière. L'inconvénient d'un tel choix est que le mode de collecte par la poste est à toute fin pratique inévitable (i.e. le numéro de téléphone est absent des informations collectées). Étant donné l'envergure du projet et aux coûts associés, les autres modes de collecte que par la poste et par téléphone n'ont pas été considérés. Les avantages de procéder par la poste sont nombreux, citons entre autres: le contrôle absolu du contenu de la communication avec les gens faisant partie de l'échantillon, la grande quantité d'information qui a pu être amassée, la facilité de stratification selon les directions régionales du Ministère et le contrôle des envois car la mise à la poste et la pose des étiquettes furent effectuées par la SAAQ. Les délais plus longs d'obtention de résultats dus à des envois postaux étaient un désavantage mineur. Bref, le choix du détenteur de permis constitue une unité échantillonnée avantageuse malgré la "non-couverture" partielle des usagers du réseau ne détenant pas de permis de conduire, tels les cyclistes, piétons, passagers non-détenteurs de permis de conduire, etc. Rappelons qu'environ 75% des résidents du Québec agés de 16 ans et plus détiennent un permis de conduire (l'âge minimum pour détenir un permis de conduire est de 16 ans au Québec). Nous avons aussi posé l'hypothèse que ce sont les conducteurs qui ont la perception la plus complète des routes. Les questionnaires utilisables ont coûté en moyenne près de 26$, ce qui nous semble raisonnable.

L'excellence de la base de sondage a été mis en lumière par le peu de détenteurs de permis non-joints ou hors de l'échantillon. Le tableau 2 présente la synthèse du nombre de détenteurs de permis selon différentes caractéristiques permettant de juger de la qualité de la base de sondage. La qualité de celle-ci est probablement due à d'efficaces procédures de mise-à-jour de la part de la SAAQ et de l'obligation légale des détenteurs de permis de signifier leur changement d'adresse à la SAAQ. (Voir tableau 2 - Annexe B)

Un autre facteur déterminant fut l'attention portée à la rédaction d'un devis descriptif détaillé. Pour obtenir satisfaction, il est impératif de bien définir ce que l'on veut. Il faut s'assurer que les objectifs et les sujets à couvrir soient très clairs. L'idée de rémunérer davantage la firme sur rendement de taux de réponse élevés fut excellente (jusqu'à concurrence de 5% du montant du contrat, à l'obtention d'un taux de réponse global supérieur à 50% et d'un minimum de 45% dans chaque région). Cette exigence a eu pour effet de forcer la firme SEGMA à imaginer une façon originale d'assurer de hauts taux de réponse. Celle-ci après autorisation de la Régie des loteries et courses du Québec a offert des prix de participation tirés au hasard parmi les détenteurs de permis sélectionnés qui avaient répondu au questionnaire dans un délai prescrit. Malgré qu'il peut paraitre contre l'éthique pour un organisme gouvernemental de s'associer à pareille démarche, d'un point de vue méthodologique, cela s'est avéré excellent. Ce problème a été atténué en faisant mention explicitement dans la lettre de SEGMA adressée aux détenteurs de permis que cette initiative était le propre de SEGMA. En tout, à chaque sondage, SEGMA a remis environ 2500$ en prix. Le fait de mentionner dans cette même lettre les probabilités de gain des prix a certainement été une bonne idée.

Une autre lettre, celle du Ministère, accompagnait chaque envoi. Celle-ci sensibilisait le répondant à l'importance de ses réponses. À chaque fois un "haut" représentant du Ministère a signé la lettre, cela renforce le niveau de sensibilisation de l'éventuel répondant. Au premier sondage, le ministre

monsieur Marc-Yvan Côté, a signé la lettre, au deuxième ce fut le sous-ministre, monsieur Jean-Marc Bard. Lors du deuxième envoi, une carte routière a été ajoutée aux questionnaire et lettres, les taux de réponse ont été meilleurs que lors du premier envoi (Voir tableau 1 - Annexe A). Ce deuxième sondage a été effectué à l'automne, un moment probablement plus propice à l'obtention de réponses en quantité.

Même si la conception du questionnaire relevait de SEGMA, il a évidemment fallu investir beaucoup d'énergies à l'interne sur celui-ci. Plusieurs unités administratives ont collaboré à cet effort, mentionnons: la Direction de la programmation, la Direction de l'entretien, la Direction des communications, la Direction de la circulation et des aménagements, le Service de la signalisation et la Direction de la recherche. Un exercice de détermination des priorités des informations à inclure au questionnaire a été mené par monsieur Louis St-Germain du Service de l'évaluation des programmes. Chaque unité administrative devait établir la priorité pour chacune des questions, ensuite cette compilation a servi de base pour arrêter le contenu final du questionnaire. La majorité des questions étaient fermées, étant donné la quantité d'information voulue. Des efforts particuliers ont ensuite été mis sur la formulation et l'agencement des questions et la présentation du questionnaire.

4. RÉSULTATS

Beaucoup de résultats ont été produits, quelques résultats synthèses du premier sondage seront présentés. Un des principaux objectifs de la démarche consistait à connaitre les différences de satisfaction des usagers du réseau routier des sujets étudiés entre le printemps et l'automne. De façon surprenante, les gens ont des niveaux de satisfaction voisins entre ces deux périodes.

Des variations significatives de satisfaction ont été observées selon les types de route. Plus les routes sont importantes, meilleure est la satisfaction. Ces variations ont surtout un lien avec les sujets suivants: la sécurité routière, l'état du réseau, la signalisation routière et évidemment l'entretien d'hiver. Ce résultat doit être interprété comme positif en même temps qu'il faut tenir compte du niveau absolu de satisfaction (par opposition au niveau relatif par type de route).

C'est le facteur humain qui représente la principale source d'insécurité. En signalisation, la visibilité des panneaux et la facilité de compréhension des messages sont évalués plus positivement, les éléments évalués négativement touchent plus spécifiquement l'éclairage, le marquage des chaussées spécifiquement la nuit et la visibilité.

L'entretien d'hiver, dans l'ensemble, est jugé bon. Les limites de vitesse satisfont la majorité des usagers. Toutefois, près du tiers des gens affirment qu'elles sont inadéquates sur les autoroutes.

Au niveau des priorités que devrait établir le Ministère, 51% des répondants préfèrent que soient privilégiés l'entretien et la réparation des routes.

5. CONCLUSION

Plusieurs facteurs déterminants ont assuré la relative réussite méthodologique de ces sondages. Le choix d'une collecte par la poste fut très judicieux. Nous pouvons affirmer que ce projet constitue un exemple à suivre dans le cadre d'une évaluation globale de services qui s'adresse à une grande population. Plusieurs organismes gouvernementaux ou grandes organisations pourraient être susceptibles de profiter d'une telle expérience. En tant que statisticien, nous avons apprécié le fait d'être consulté dès le début du projet.

RÉFÉRENCES

Fluet, Michel, Sam Inc., Formalisation du calcul des taux de réponse en présence d'unités hors population, Recueil des textes des présentations du colloque sur les applications de la statistique, B.S.Q., mai 1987.

Gouvernement du Québec, ministère des Transports, Sondage d'opinion auprès des usagers du réseau routier, résumé des résultats, mars 1980.

Lavalin, Segma (86) Inc., Sondage d'opinion pour évaluer la satisfaction des usagers du réseau routier, septembre 1989.

Lavalin, Segma (86) Inc., Sondage d'opinion pour évaluer la satisfaction des usagers du réseau routier (phase 2), mars 1990.

Statistique Canada, Conception des questionnaires, Manuel d'atelier, troisième édition, août 1981.

Source d'information: Paul Bergeron, ministère des Transports, Direction de la recherche, Service de la statistique et de l'économie des transports, 700, boul. St-Cyrille est, 17e étage, Québec, (Québec) G1R 5H1 (418) 644-1492

PROJET PILOTE D'UN CENTRE DE MICROGRAPHIE

PRESENTATION FAITE AU COLLOQUE SUR
LES APPLICATIONS DE LA STATISTIQUE
DANS LE CADRE DU 58E CONGRES DE
L'ASSOCIATION CANADIENNE FRANCAISE POUR
L'AVANCEMENT DES SCIENCES

LE 14 MAI 1990
A L'UNIVERSITE LAVAL

PIERRE CORRIVEAU
SERVICE DES RECHERCHES STATISTIQUES
MINISTERE DU REVENU DU QUEBEC

1. INTRODUCTION

2. ECHANTILLON EXPLORATOIRE

2.1 Objectifs

2.2 Méthodologie d'échantillonnage

2.3 Résultats

3. ECHANTILLON DANS UN CONTEXTE DE RODAGE

3.1 Objectifs

3.2 Méthodologie d'échantillonnage

3.3 Résultats

4. ECHANTILLON DANS UN CONTEXTE DE PRODUCTION

4.1 Objectif

4.2 Méthodologie d'échantillonnage

4.3 Résultats

5. CONCLUSION

1. INTRODUCTION

Les problèmes rencontrés actuellement avec l'utilisation des dossiers physiques sont les coûts élevés de l'entreposage et des ressources pour en supporter la gestion, l'accès difficile et les délais occasionnés dans les communications avec les usagers. Il est question ici de 16 millions de déclarations environ, 4 millions/année pour 4 années d'imposition.

Le Ministère du Revenu a mandaté une équipe d'une étude sur la rentabilité de l'implantation de centres de micrographie à ses bureaux de Montréal et Québec. Un projet-pilote a donc été mis sur pied au bureau de Montréal.

La micrographie consiste en la conversion de documents sur microfilms avec indexation informatisée des dossiers microfilmés. Elle implique le développement, la duplication et le contrôle de la qualité des films produits. Le repérage des dossiers est assisté par ordinateur et leur diffusion s'effectue par fac-similé.

Des échantillons statistiques de déclarations de revenus de particuliers ont été utilisés lors des différentes activités de ce projet.

La firme Bell & Howell a participé au projet en tant que partenaire technologique. Elle avait le mandat de fournir le support technologique nécessaire aux opérations d'un centre de micrographie autant pour les logiciels informatiques que pour l'équipement.

2. ECHANTILLON EXPLORATOIRE

2.1 Objectifs

Les objectifs sont, d'abord, de connaître la population documentaire en ce qui concerne les caractéristiques physiques des documents que l'on retrouve dans les dossiers et, ensuite, de constituer un jeu standard de dossiers afin de tester et choisir l'équipement approprié.

2.2 Méthodologie d'échantillonnage

Les déclarations cotisées informatiquement au bureau de Montréal pour l'année 1986 ont servi de base d'échantillonnage. Le choix des déclarations du bureau de Montréal s'explique par une raison d'ordre pratique, le projet se tenait à ce bureau.

L'unité d'échantillonnage était la déclaration des particuliers. Les déclarations ont été sélectionnées systématiquement à l'intérieur des strates qui ont été formées à partir du type de déclarations complétées et des annexes jointes. Un échantillon de 1 202 dossiers a ainsi été sélectionné et 1 163 de ceux-ci ont été traités, 39 étant non-disponibles.

Les principales informations recueillies lors du traitement des dossiers sont :

- Le nombre de documents dans le dossier. Un document étant toutes pièces présentes, à partir de l'enveloppe elle-même, jusqu'au dernier reçu joint à la déclaration.

et pour chacun des documents présents

- l'identification
- le nombre de copies identiques
- le nombre d'images
- le type
- la condition générale
- les dimensions
- l'épaisseur
- la couleur de fond
- la couleur d'impression
- la présence ou non de bretelles

2.3 Résultats

A partir de l'échantillon on estime à 25 843 303 le nombre de documents et à 31 484 150 le nombre d'images contenus dans la population des 2 413 485 déclarations de la population observée.

Parmi l'ensemble des caractéristiques recueillies, la plus importante touchait la dimension des documents. La proportion estimée de documents de dimension standard, de hauteur comprise entre 5 et 14 pouces inclusivement et de largeur comprise entre 4,5 et 11 pouces inclusivement, est 50,6 % seulement. Ce résultat a nécessité la création d'un mécanisme d'alimentation spécial adapté aux documents de petit format pour les caméras qui servent à la prise d'image.

3. ECHANTILLON DANS UN CONTEXTE DE RODAGE

3.1 Objectifs

Les objectifs sont, premièrement, de roder les équipements et les procédures, deuxièmement, d'atténuer l'effet d'apprentissage et, finalement, de fournir de l'information sur la distribution des caractéristiques à l'étude, soit le temps de préparation des dossiers et celui de la prise d'image.

3.2 Méthodologie d'échantillonage

Une liste des numéros de lots créée à partir du numéro de lot des déclarations cotisées informatiquement au bureau de Montréal pour l'année 1987 saisies sur la Banque Centrale des Données (BCD) a servi de base d'échantillonnage.

L'unité d'échantillonnage était le lot de déclarations. Un lot consiste en un regroupement de déclarations selon leur type et peut en contenir jusqu'à 100. Les lots ont été sélectionnés de façon systématique à l'intérieur de chacune des strates. Les lots étaient triés selon le nombre de déclarations contenues dans ceux-ci. Les strates ont été créées selon trois (3) critères. Premièrement selon le type de déclarations contenues dans le lot: abrégées, détaillées, sans calcul ou mélangées, deuxièmement selon que les déclarations sont des originales, des corrigées ou des "alpha" et, troisièmement selon le code alphabétique du numéro de lot qui est assigné séquentiellement lors de la création des lots. Un échantillon de 226 lots a ainsi été sélectionné pour un total de 10 267 déclarations.

3.3 Résultats

Cet échantillon a permis d'ajuster et de roder les divers équipements et procédures et a permis d'atténuer l'effet d'apprentissage pour les mesures du prochain échantillon. Il n'a pu fournir d'estimation de la distribution des données.

4. ECHANTILLON DANS UN CONTEXTE DE PRODUCTION

4.1 Objectif

L'objectif de cet échantillon était de simuler les opérations du centre dans un contexte de production avec spécialisation des tâches des employés afin d'estimer les ressources nécessaires pour la préparation et le microfilmage de la population de déclarations cotisées informatiquement au bureau de Montréal pour l'année d'imposition 1987.

4.2 Méthodologie d'échantillonnage

La méthodologie utilisée est la même que celle de l'échantillon dans un contexte de rodage à l'exception du pas de sondage. Il a été légèrement augmenté afin de diminuer quelque peu la taille de l'échantillon.

Le nombre de lots sélectionnés a été de 185 pour un total de 8 164 déclarations, 175 de ces lots ont été reçus pour 8 076 déclarations traitées.

4.3 Résultats

Les estimations des ressources nécessaires sont de 22 505 jours/personne pour la préparation des dossiers et de 8 127 jours/personne pour le microfilmage. Ces estimations ont été calculées à partir des temps observés lors du traitement des dossiers. Le départ de quelques employés remplacés par des nouveaux a fait réapparaître l'influence de la période d'apprentissage dans le niveau de production. Un indice de rendement associé à chacun des employés a été calculé afin de corriger cette courbe d'apprentissage.

Les nouvelles estimations calculées à partir des temps ajustés sont de 16 280 jours/personne pour la préparation et 8 716 jours/personne pour le micorfilmage de la population de déclarations.

L'estimation du nombre moyen d'images par dossier est de 14,5.

5. CONCLUSION

Les résultats du projet-pilote d'un centre de micrographie ont clairement démontré qu'il serait très dispendieux et non-rentable de microfilmer intégralement l'ensemble des dossiers avec un processus de haute qualité et de haute précision.

Afin d'en arriver à une solution moins coûteuse, une approche alternative a été envisagée, basée sur certaines hypothèses.

Considérant qu'environ seulement trente pour cent (30 %) des dossiers font l'objet de requête ultérieure à la cotisation initiale, il pourrait être possible de discriminer, à priori, la population de dossiers. Suite à cette discrimination, un dossier à faible probabilité de requête pourrait être épuré avant d'être microfilmé ou, tout simplement, détruit.

Un mandat est présentement en cours dans le but de discriminer les dossiers en catégories, lesquelles seraient traitées de façon différente. Cette approche permettrait de rentabiliser à moyen terme l'implantation des centres de micrographie.

UNE APPROCHE POUR L'ÉCHANTILLONNAGE ET L'ESTIMATION D'ENQUÊTES TRIMESTRIELLES: l'Enquête sur le camionnage pour compte d'autrui[1]

On présente ici une méthode simple pour l'échantillonnage d'enquêtes trimestrielles où les échantillons sont répartis annuellement pour les quatres trimestres. De plus, on présentera les estimateurs de totaux et de tendances entre les trimestres, et, pour chacun d'eux, l'estimateur de la variance. Enfin, on présentera un exemple de la méthode proposée: l'Enquête trimestrielle sur le camionnage pour compte d'autrui.

MOTS-CLÉS: enquête trimestrielle, répartition exponentielle, estimation de covariances, plan à deux degrés.

1. INTRODUCTION

L'Enquête annuelle sur le camionnage pour compte d'autrui existe depuis 1972. Afin de mieux mesurer l'impact de la déréglementation du transport routier en vigueur depuis 1988, et de répondre plus rapidement aux besoins des usagers de l'enquête, il a été décidé de transformer l'enquête annuelle en enquête trimestrielle.

2. L'ENQUÊTE ANNUELLE SUR LE CAMIONNAGE POUR COMPTE D'AUTRUI

Le but de l'Enquête annuelle sur le camionnage pour compte d'autrui (EACCA) est d'obtenir des estimations sur les livraisons, les recettes, les tonnes et les tonnes-kilomètres des compagnies de camionnage oeuvrant pour le compte d'autrui.

On désire produire des estimations par groupe de marchandises et par paire origine/destination.

La population cible de l'EACCA comprend toutes les compagnies de camionnage pour compte d'autrui avec un revenu interurbain annuel d'au moins 100,000$. Pour l'enquête trimestrielle, on a haussé cette borne inférieure à 500,000$.

On utilise un plan de sondage à deux degrés avec échantillonnage aléatoire simple stratifié au ler degré et systématique au 2ème degré. L'unité primaire est une compagnie de camionnage dans le champ de l'enquête. L'unité secondaire est une livraison interurbaine effectuée par cette compagnie.

Au premier degré, les unités sont stratifiées par classe de revenu interurbain annuel, groupe de marchandise et région d'exploitation. La classe de revenu 1 est autoreprésentative. Il n'y a cependant pas de stratification au deuxième degré.

---

1 Jean Dumais, Pierre Lavallée, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa, K1A 0T6

Pour plus de détails sur l'EACCA, on peux consulter Dumais et Leroux (1987).

3. EXIGENCES DE BASE POUR L'ENQUÊTE TRIMESTRIELLE

La transformation de l'enquête annuelle en enquête trimestrielle devait se faire sous certaines contraintes visant à minimiser l'impact sur les séries de données et les opérations.

D'abord, l'Enquête trimestrielle sur le camionnage pour compte d'autrui (ETCCA) devait utiliser une seule base de sondage annuelle, c'est-à-dire sans mise à jour trimestrielle. Comme dans le cas de l'EACCA, on devait considérer les compagnies importantes comme autoreprésentatives. Pour l'enquête trimestrielle, une unité autoreprésentative est échantillonnée avec certitude durant les quatre trimestres de l'année. Afin de réduire le fardeau de réponse, les unités des strates non-autoreprésentatives ne pouvaient être visitées qu'une seule fois durant l'année; ce qui demande une sélection d'un minimum de 4 x 2 compagnies par strate pour obtenir une estimation sans biais de la variance des estimations. On a donc considéré les strates possédant huit compagnies ou moins comme autoreprésentatives. Finalement, pour pouvoir maintenir la charge de travail constante durant toute l'année, on devait répartir l'échantillon également entre les trimestres.

Ces contraintes sont décrites en détails par Lavallée (1989a).

4. RÉPARTITION ET SÉLECTION DE L'ÉCHANTILLON

4.1 CALCUL DE LA TAILLE D'ÉCHANTILLON

Posons la taille d'échantillon annuelle n^A fixée à l'avance.

D'abord, le nombre total d'unités autoreprésentatives est:

$$n^{A,AR} = \sum_{h \in AR} n_h^{A,AR} = 4 \sum_{h \in AR} N_h$$

où N_h est le nombre d'unités primaires de la strate h.

Ensuite, on répartit les $n^{A,NAR} = n^A - n^{A,AR}$ unités restantes aux strates non-autoreprésentatives.

$$n_h^{A,NAR} = a_h n^{A,NAR}$$

où

$$a_h = \frac{N_h^p S_h^q N_h^r}{\sum_{h \in NAR} N_h^p S_h^q N_h^r} ,$$

$p \geq 0, q \geq 0, r \geq 0$.

$$X_h = \sum_{i=1}^{N_h} x_{hi} \ .$$

$$S_h = \sqrt{\frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left(x_{hi} - \frac{X_h}{N_h} \right)^2} \ .$$

On utilise comme variable "x" le revenu interurbain annuel. Cette méthode de répartition, dite exponentielle, est dérivée de la méthode proposée par Bankier (1988). On choisit une méthode en particulier en posant les valeurs des paramètres p, q et r. Par exemple, avec p=1, q=1 et r=0, on obtient la répartition de Neyman, décrite dans Cochran (1977).

En rapport avec les contraintes décrites précédemment, on s'assure en pratique que $n_h^{A,NAR} \geq 8$ pour chaque strate non-autoreprésentative. Enfin, on obtient la taille des échantillons non-autoreprésentatifs trimestriels n_h^Q en divisant $n_h^{A,NAR}$ par 4.

Pour plus de détails, on peut consulter Lavallée (1989b).

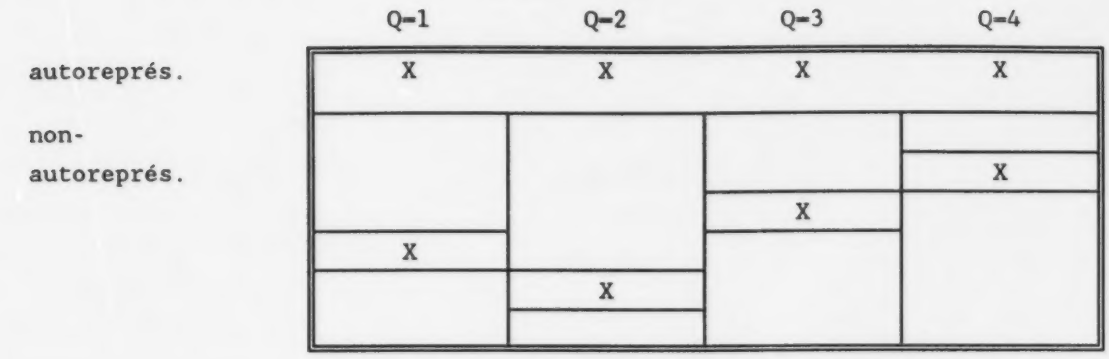
4.2 SÉLECTION DES UNITÉS PRIMAIRES ET SECONDAIRES

Pour la sélection aléatoire des unités primaires (ou compagnies) pour les strates non-autoreprésentatives, on procède quatre fois à des tirages aléatoires simples sans remise de n_h^Q unités. Ainsi, on tire n_h^1 unités primaires parmi N_h pour le trimestre Q=1, n_h^2 unités parmi $N_h - n_h^1$ pour le trimestre Q=2, et ainsi de suite pour les trimestres Q=3 et Q=4.

On effectue la sélection trimestrielle des unités secondaires par échantillonnage systématique à partir de chaque unité primaire (ou compagnie) choisie. La fraction de sondage utilisée dépend du nombre total de livraisons interurbaines effectuées par la compagnie au cours du trimestre. En pratique, on utilise une table de fractions de sondages telle que celle donnée en annexe A.

Le diagramme 1 ci-dessous donne une idée de l'organisation de l'échantillon. A chaque trimestre, on collecte les renseignements auprès d'une partie autoreprésentative et d'une partie non-autoreprésentative de la population.

Diagramme 1



5. AVANTAGES ET DÉSAVANTAGES

L'approche proposée offre des avantages certains au prix de quelques inconvénients par rapport à un plan de sondage avec un chevauchement entre les trimestres au niveau des strates non-autoreprésentatives et avec la collecte de données effectuée pour plusieurs trimestres à la fois. Ce dernier plan de sondage est souvent considéré pour des enquêtes infra-annuelles.

Au nombre des avantages, notons la simplicité d'implantation, la possibilité de produire des estimations finales après chaque trimestre, la réduction du fardeau de réponse des compagnies non-autoreprésentatives, et le passage simple d'une année à l'autre par l'utilisation d'échantillons indépendants.

Le principal désavantage de cette approche est la plus grande variabilité entre les estimations trimestrielles (estimation des tendances) par manque de chevauchement des échantillons successifs.

6. ESTIMATION

Pour simplifier, nous présentons la solution pour un plan à un degré. Les résultats peuvent être facilement étendus à un plan à deux degrés. Ces derniers résultats sont décrits dans Lavallée (1990).

6.1 ESTIMATION DE TOTAUX TRIMESTRIELS

Pour un trimestre donné Q, l'échantillon est un échantillon aléatoire simple stratifié de n_h^Q compagnies choisies parmi N_h.

On estime le total trimestriel Y_Q par

$$\hat{Y}_Q = \sum_h N_h \bar{y}_{Qh}$$

où

$$\bar{y}_{Qh} = \frac{1}{n_h^Q} \sum_{i=1}^{n_h^Q} y_{Qhi}.$$

La variance $Var(\hat{Y}_Q)$ peut être estimée par

$$var(\hat{Y}_Q) = \sum_h N_h^2 \left(\frac{1}{n_h^Q} - \frac{1}{N_h} \right) s_{Qh}^2$$

où

$$s_{Qh}^2 = \frac{1}{n_h^Q - 1} \sum_{i=1}^{n_h^Q} (y_{Qhi} - \bar{y}_{Qh})^2 .$$

6.2 ESTIMATION DE TOTAUX ANNUELS

On estime le total annuel Y_A par

$$\hat{Y}_A = \sum_{Q=1}^{4} \hat{Y}_Q .$$

De par la méthode de répartition, les échantillons trimestriels ne sont pas mutuellement indépendants.

Alors,

$$Var(\hat{Y}_A) = \sum_{Q=1}^{4} Var(\hat{Y}_Q) + 2 \sum_{Q=1}^{4} \sum_{Q'<Q} Cov(\hat{Y}_Q, \hat{Y}_{Q'})$$

$$\neq \sum_{Q=1}^{4} Var(\hat{Y}_Q)$$

puisque, d'après Tam (1984), $Cov(\hat{Y}_Q, \hat{Y}_{Q'}) = - \sum_{h \in NAR} N_h S_{QQ'h} \neq 0$.

Le terme de covariance $S_{QQ'h}$ peut être estimé par

$$s_{QQ'h} = \frac{1}{n_h^Q + n_h^{Q'} - 1} \left\{ \frac{\bar{y}_{Q'h}}{\bar{y}_{Qh}} (n_h^Q - 1) s_{Qh}^2 + \frac{\bar{y}_{Qh}}{\bar{y}_{Q'h}} (n_h^{Q'} - 1) s_{Q'h}^2 \right\}.$$

Notons que $s_{QQ'h}$ est obtenu en imputant les données des compagnies absentes du trimestre Q (ou Q') à partir de celles présentes dans le trimestre complémentaire Q' (ou Q) (Voir le Diagramme 1 ci-haut.). La méthode d'imputation utilisée est celle dite "historique par tendance". Elle consiste à estimer à l'aide d'un estimateur par quotient une donnée absente d'un trimestre avec, comme variable auxiliaire, la même donnée présente pour un autre trimestre.

Ainsi, pour chaque strate non-autoreprésentative:

Q	Q'
$y_{Qh1}, \ldots, y_{Qhn}$	$\hat{y}_{Q'h1}, \ldots, \hat{y}_{Q'hn}$
$\hat{y}_{Qh1}, \ldots, \hat{y}_{Qhm}$	$y_{Q'h1}, \ldots, y_{Q'hm}$

où $\hat{y}_{Qhi} = \frac{\bar{y}_{Qh}}{\bar{y}_{Q'h}} y_{Q'hi}$ et $\hat{y}_{Q'hi} = \frac{\bar{y}_{Q'h}}{\bar{y}_{Qh}} y_{Qhi}$.

6.3 ESTIMATION DE TENDANCES TRIMESTRIELLES

La réalisation d'une enquête trimestrielle permet l'estimation des tendances entre occurrences successives. Ainsi on désire estimer $T_{QQ'}$ où $Q' < Q$. Il vient naturellement que

$$\hat{T}_{QQ'} = \frac{\hat{Y}_Q - \hat{Y}_{Q'}}{\hat{Y}_{Q'}}$$

$$= \frac{\hat{Y}_Q}{\hat{Y}_{Q'}} - 1$$

$$= \hat{R}_{QQ'} - 1.$$

L'estimateur $\hat{T}_{QQ'}$ est un estimateur par quotient combiné (voir Cochran (1977).).

On peut estimer sa variance $Var(\hat{T}_{QQ'})$ par

$$var(\hat{T}_{QQ'}) = \frac{1}{\hat{Y}^2_{Q'}} var(\hat{Y}_Q - \hat{R}_{QQ'}\hat{Y}_{Q'})$$

$$= \frac{1}{\hat{Y}^2_{Q'}} \{var(\hat{Y}_Q) + \hat{R}^2_{QQ'} var(\hat{Y}_{Q'}) - 2\hat{R}_{QQ'} cov(\hat{Y}_Q, \hat{Y}_{Q'})\}.$$

7. CONCLUSION

Nous avons présenter une méthode simple pour l'échantillonnage et l'estimation d'enquêtes trimestrielles. Cette méthode satisfait les exigences suivantes:

- utiliser une base de sondage annuelle;
- considérer les grosses compagnies comme autoreprésentatives;
- sélectionner les petites compagnies une seule fois durant l'année;
- répartir l'échantillon également entre les trimestres.

La méthode exposée ici permet d'obtenir des estimations finales après chaque trimestre par rapport à un plan de sondage où il y a chevauchement entre les trimestres et où la collecte de données effectuée pour plusieurs trimestres à la fois.

8. BIBLIOGRAPHIE

Cochran, W.G. (1977), "Sampling Techniques", 3ème édition, John Wiley and Sons, New York, 1977.

Bankier, M.D. (1988), "Power Allocations: Determining Sample Sizes for Subnational Areas", The American Statistician, vol. 42, no. 3, août 1988.

Dumais, J., Leroux, D. (1987), "Enquête annuelle sur l'origine et destination des marchandises - Méthodologie pour la stratification de l'univers et l'allocation de l'échantillon", Cahier de travail no. BSMD-87-019F, décembre 1987.

Lavallée, P. (1989a), "Plan de sondage de l'enquête trimestrielle sur le camionnage pour compte d'autrui", Rapport de la Division des méthodes d'enquêtes-entreprises, septembre 1989.

Lavallée, P. (1989b), "Quarterly For-Hire Trucking Survey Sample Allocation and Selection System, Statement of Requirements (Version 2)", Rapport de la Division des méthodes d'enquêtes-entreprises, octobre 1989.

Lavallée, P. (1990), "Quarterly For-Hire Trucking Survey Weighting and Estimation System, Statement of Requirements (Version 2)", Rapport de la Division des méthodes d'enquêtes-entreprises, février 1990.

Tam, S.M. (1984), "On Covariances From Overlapping Samples", The American Statistician, vol. 38, no. 4, 1984.

ANNEXE A

Table des fractions de sondages pour les unités secondaires

Nombre total de livraisons interurbaines dans le trimestre			Intervalle	Nombre minimum de livraisons sélectionnées	Nombre maximum de livraisons sélectionnées	Départ aléatoire
0	-	150	1	0	150	1
151	-	300	2	75	150	1
301	-	600	4	75	150	3
601	-	1,500	8	75	188	7
1,501	-	2,500	12	125	209	9
2,501	-	5,000	20	125	250	14
5,001	-	7,500	30	167	250	23
7,501	-	10,000	40	187	250	29
10,001	-	15,000	50	200	300	35
15,001	-	22,500	70	214	322	18
22,501	-	35,000	100	225	350	23
35,001	-	60,000	150	233	400	111
60,001	-	100,000	200	300	500	121
100,001	-	200,000	300	333	667	44
200,001	-	600,000	400	500	1500	350
600,001	-	900,000	500	1200	1800	8
900,001	-	1,200,000	650	1384	1847	392
1,200,001	-	1,600,000	850	1412	1883	788
1,600,001	-	+∞	2000	800	+∞	1399

UN LOGICIEL D'AIDE À LA SÉLECTION DES PLANS D'ÉCHANTILLONNAGE EN CONTRÔLE DE LA QUALITÉ

Tony Labillois
Statistique Canada
Division des méthodes d'enquêtes - Entreprises
10-J, Immeuble R.-H. Coats
Ottawa, K1A 0T6

Résumé

A Statistique Canada, le contrôle statistique de la qualité est employé pour vérifier les étapes de traitement des données d'enquêtes réalisées dans divers secteurs d'activité. Il faut donc fréquemment employer et développer des méthodes et des plans d'échantillonnage spécifiques suivant divers facteurs. Dans cette tâche, l'informatique peut aider grandement le méthodologiste en lui évitant de nombreux calculs. Après avoir situé le contexte de travail, on présentera un logiciel d'aide à la détermination et à l'évaluation des plans simples de Dodge-Romig avec limite de qualité moyenne à la sortie (AOQL) qui assurent une inspection minimale pour une certaine qualité à l'entrée.

Mots clés: contrôle statistique de la qualité, contrôle par échantillonnage par attributs, limite de qualité moyenne à la sortie (AOQL), échantillonnage passe-lots.

Introduction

Depuis quelques années, on peut constater l'apparition d'une prise de conscience collective relativement à l'importance de la qualité des produits et des services disponibles sur le marché. Les consommateurs exigent que les produits respectent des normes plus strictes de fiabilité pour une meilleure efficacité tandis que les fournisseurs comprennent qu'en améliorant leur production ils réalisent des économies substantielles. Statistique Canada, à titre de fournisseur de renseignements statistiques, n'a pas échappé à cette tendance. Grâce au développement de systèmes généralisés pour la réalisation des enquêtes, aux programmes d'assurance et de contrôle de la qualité des données et à d'autres initiatives, l'agence gouvernementale peut continuer à fournir des résultats de plus en plus fiables malgré les contraintes budgétaires actuelles.

Dans cet effort de standardisation des méthodes pour une qualité accrue, la section des méthodes d'assurance de la qualité a elle aussi un rôle déterminant à jouer. Déjà, certains systèmes ont été conçus pour aider à la généralisation et à l'automatisation des opérations de contrôle de la qualité (CQ) des données. Maintenant, les méthodologistes de cette section visent également à standardiser ou à informatiser davantage les étapes de développement des nouvelles opérations de CQ. Un bel exemple de cette dernière tendance se situe au niveau de l'étape de sélection des plans d'échantillonnage pour le contrôle statistique des lots. En effet, au cours de 1988-89, on a développé une méthodologie de standardisation des plans d'échantillonnage qui se reflète aujourd'hui dans un ensemble de plans types qui peuvent être employés pour de nombreuses opérations [1]. Cependant, il n'en demeure pas moins que des plans doivent être calculés spécifiquement pour certains cas où les plans types ne seraient pas nécessairement appropriés. C'est pour cette raison, pour simplifier les calculs et pour augmenter la rapidité que la réalisation d'un logiciel de calcul et d'évaluation de plans d'échantillonnage a été entreprise.

Après avoir discuté plus en détail du rôle et des objectifs de la section des méthodes d'assurance de la qualité et du contexte global de travail, le logiciel d'aide à la sélection des plans d'échantillonnage et ses principaux modules actuels seront décrits. On parlera également des ajouts envisagés pour le futur.

1. Rôle et objectifs de la Section des méthodes d'assurance de la qualité

La section des méthodes d'assurance de la qualité (SMAQ) de la Division des méthodes d'enquêtes - Entreprises (DMEE) s'occupe d'assurance et de contrôle de la qualité dans de nombreuses enquêtes réalisées par Statistique Canada. Rappelons que l'assurance de la qualité concerne l'ensemble de toutes les activités qui donnent l'évidence nécessaire pour assurer qu'un produit statistique est de qualité suffisante pour satisfaire les besoins. Le contrôle de la qualité s'applique plutôt aux étapes de traitement des données pour assurer que les erreurs introduites par les préposés au traitement n'excèdent pas le niveau déterminé [2]. Le rôle du groupe de CQ est de

> «...promouvoir, développer, implanter et maintenir les méthodes de contrôle de la qualité pour les opérations de traitement des enquêtes de Statistique Canada lorsque ces méthodes sont applicables.» [3]

En général, la SMAQ est consultée par les autres divisions de Statistique Canada qui désirent effectuer un contrôle de la qualité de leurs données après l'une ou plusieurs des étapes de traitement (codage/transcription, vérification, saisie, etc). La section donne des conseils pour l'élaboration d'un programme convenable de contrôle de la qualité compte tenu des besoins (niveau de qualité avant et après l'opération), du temps et des ressources disponibles.

2. Stratégie globale de contrôle statistique de la qualité

Pour arriver à obtenir des résultats qui correspondent aux attentes des divisions clientes, la SMAQ a adopté une stratégie globale pour le contrôle statistique de la qualité qui est appliquée à un certain niveau à la majorité des opérations pour lesquelles la section a été consultée. Cette stratégie comprend plusieurs étapes de l'approche d'«Acceptance Control» proposé par Edward Schilling [4] et est schématisée à l'annexe 1.

Dans le traitement des données, le travail des préposés est souvent très répétitif et c'est pour cela qu'il s'y introduit des erreurs qui proviennent de sources identifiables (manque de formation, fatigue, matériel non adéquat) et aussi de sources non identifiables. Même si certaines causes sont identifiées, cela ne veut pas dire pour autant que l'on puisse corriger la situation. De plus, lorsqu'un préposé débute le traitement d'une nouvelle enquête, il fait beaucoup d'erreurs dues au manque de pratique. Ce n'est qu'après un certain temps qu'il peut espérer atteindre une variabilité plus constante (préposé stable). Il ne faut pas oublier que, même s'ils connaissent bien les procédures de traitement, certains préposés n'atteindront jamais le niveau de la stabilité. Chaque préposé est donc considéré comme étant un procédé unique qui doit être contrôlé individuellement. C'est, entre autres, pour ces raisons que le contrôle statistique par échantillonnage des lots occupe une place importante à la base de la stratégie globale.

Lorsqu'un préposé débute un nouveau traitement, on vérifie généralement ses premiers lots à 100% pour obtenir son niveau de qualité (taux d'erreurs) à l'entrée du contrôle. On peut ensuite lui attribuer un plan d'échantillonnage (simple de Dodge-Romig avec AOQL [5]) approprié selon sa performance. La mesure de sa performance qui sert à la qualification est connue comme étant la moyenne de traitement [6] et est calculée (avec d'autres statistiques) par le Système de traitement des données du contrôle de la qualité (STDCQ) [7] pour chaque préposé à la fin de chaque période de traitement. Ensuite, à mesure que l'on note une amélioration (ou une détérioration) de sa performance à chaque période de traitement, on lui attribue un nouveau plan d'échantillonnage moins (ou plus) sévère avec un niveau d'inspection moins (ou plus) élevé. Si le préposé devient bon ou très bon dans son travail, on lui permet de passer aux plans d'échantillonnage passe-lots [8]. Rappelons brièvement que ceci consiste à vérifier par échantillonnage tous les lots jusqu'à ce que i lots successifs soient acceptés (généralement i=4). Ensuite on vérifie par échantillonnage 1 lot à tous les f lots (f augmente si le préposé s'améliore) jusqu'à ce que l'on rejette un lot. A ce moment-là, le travail du préposé retourne à une inspection normale par échantillonnage jusqu'à ce qu'il

obtienne à nouveau i lots successifs acceptés. On peut voir que le niveau d'inspection diminue beaucoup dans cette étape tout en conservant une bonne qualité. Jusqu'à maintenant, on peut remarquer que la rectification des erreurs par les préposés est présente à chaque étape ainsi qu'un système de rétroaction qui informe ies différents intervenants de la performance de l'opération pour la période précédente. Ces deux facteurs améliorent la qualité ainsi que l'intérêt des intervenants envers celle-ci.

Un projet de recherche est présentement en cours pour implanter le contrôle de procédé et les cartes de contrôle aux préposés excellents et stables. Ceci permettrait de réduire encore davantage le niveau d'inspection et par conséquent amènerait des économies importantes tout en maintenant une bonne qualité.

Tout au long de ces étapes un objectif principal est visé: maintenir et améliorer le niveau de qualité du produit tout en réduisant continuellement le niveau d'inspection et par conséquent les coûts associés à l'opération.

On peut donc constater dans cette stratégie l'importance des plans d'échantillonnage qui sont à la base de son bon fonctionnement et qui varient pour chaque préposé et pour chaque opération de traitement. C'est à cause de cette importance que l'on a conçu un ensemble de plans types et un logiciel d'aide à la sélection.

3. Logiciel d'aide à la sélection des plans d'échantillonnage (PLANSCQ/QCPLANS)

En septembre 1989, après la réalisation des plans types, il fut décidé qu'un logiciel de calcul et de sélection des plans serait un outil important puisque les plans types ne sont pas applicables partout et que le fardeau de calcul des plans augmente avec le nombre d'opérations qui ne peuvent pas employer ces plans standardisés. Il existait déjà certains des modules qui composent aujourd'hui le logiciel mais ils étaient éparpillés et leur accès nécessitait des procédures assez longues. Il fut convenu que le développement débuterait immédiatement en intégrant les programmes déjà existants et en concevant les modules manquants pour répondre adéquatement aux besoins des méthodologistes. Maintenant, un logiciel facile d'accès, avec des écrans d'assistance et contrôlé par menus est disponible sur les micro-ordinateurs de la SMAQ. Les résultats des modules peuvent également être imprimés aisément au moyen d'une imprimante reliée au micro-ordinateur. La partie interactive de ce logiciel est bilingue (français et anglais). Les modules de ce logiciel sont écrits en GWBASIC, en FORTRAN, en SAS et en APL et peuvent fonctionner sur un micro-ordinateur compatible au IBM PC. Voici maintenant une description des principaux modules actuels de ce logiciel.

3.1. PLANCALC : Calculateur de plans d'échantillonnage

Ce module [9] a pour but de calculer les plans d'échantillonnage simples de Dodge-Romig avec l'indice de qualité AOQL et contrôle minimal dans le cas de lots dont l'effectif est fixe ou variable, assumant que la qualité des lots est conforme aux distributions de Poisson ou hypergéométriques.

3.1.1. Lots dont l'effectif est fixe

Ce module demande à l'utilisateur l'effectif N des lots, l'AOQL recherché et le taux d'erreurs d'entrée hypothétique p (niveau de qualité avant le contrôle). A l'aide de ces données, il calcule l'effectif n des échantillons à prélever, le nombre d'acceptation c (nombre maximal d'erreurs acceptables dans un lot; s'il y a plus de c erreurs, on rejette le lot), la probabilité d'accepter le lot Pr(accept.), l'AOQ (qualité moyenne après contrôle), l'AOQL exact et le taux de contrôle probable EIR (niveau d'inspection).

Si la solution optimale correspond à un plan avec $c=0$, il calcule aussi deux autres plans à envisager, dans lesquels $c=1$ parce que la courbe d'efficacité des plans avec $c=0$ n'a pas de bonnes propriétés et qu'ils découragent les préposés en semblant demander la perfection.

3.1.2. Lots dont l'effectif est variable

Pour produire un ensemble de plans selon la taille variable du lot, le module demande à l'utilisateur l'effectif maximal N des lots, la taille minimale n de l'échantillon à prélever, l'AOQL choisi et le taux d'erreurs d'entrée hypothétique p. Il demande aussi l'écart maximal admissible entre l'effectif des échantillons calculé pour le lot le plus petit et celui calculé pour le lot le plus grand. En outre, l'utilisateur peut autoriser ou interdire les plans sans éléments défectueux (c.-à-d. c=0) dans l'ensemble des plans calculés.

Les éléments suivants sont calculés pour chaque intervalle d'effectifs de lots : l'effectif n des échantillons, le nombre d'acceptation c, Pr(accept.), l'intervalle des AOQ, l'intervalle des AOQL exacts et l'intervalle des EIR.

3.2. PLANALYZER : Analyseur de plans d'échantillonnage

Ce module [9] a pour but d'évaluer des plans d'échantillonnage déjà calculés en fonction de leurs caractéristiques d'efficacité. Rappelons que la courbe d'efficacité est le graphique qui montre la probabilité d'acceptation des lots Pr(accept.) en fonction du niveau d'erreur à l'entrée p. Le module fonctionne pour $0 <= c <= 9$ et utilise des distributions de probabilité de Poisson et hypergéométriques.

Le module demande à l'utilisateur l'effectif N des lots, l'effectif n des échantillons et le nombre d'acceptation c de n'importe quel plan d'échantillonnage de référence, ainsi que l'intervalle de validation i et la fréquence d'échantillonnage f des lots dans le cas des plans d'échantillonnage passe-lots. Il permet à l'utilisateur de spécifier les taux d'erreurs d'entrée (c.-à-d. les valeurs de p) à utiliser dans l'évaluation, d'utiliser les valeurs de p spécifiées dans le passage précédent de la même session ou de sélectionner les valeurs implicites proposées par le module. Il calcule les données suivantes, premièrement pour le plan d'échantillonnage de référence, puis éventuellement pour le plan d'échantillonnage passe-lots : Pr(accept.), EIR, AOQ et ARL (mesure du nombre probable de lots successifs qui devraient être considérés sous contrôle) pour chacune des 12 valeurs spécifiées pour p. En outre, l'AOQL est calculé ainsi que la valeur correspondante de p.

3.3. PLANMOD : Programme d'évaluation des effets des modifications aux plans d'échantillonnage

Le programme [9] permet d'examiner rapidement un maximum de 35 plans d'échantillonnage proches d'un plan de base calculé par le calculateur de plans. Ceci permet d'explorer des solutions proches de la solution optimale pour respecter certaines contraintes qui sont présentes dans certains cas. Par exemple, pour des raisons économiques, on peut tenter de trouver un plan qui nécessiterait moins de contrôle même si on doit obtenir une qualité légèrement moins bonne (un plus grand AOQL). En outre, il est possible d'étudier un maximum de 8 plans supplémentaires optimaux, dans lesquels les nombres d'acceptation sont proches du nombre d'acceptation du plan de base.

Le module demande à l'utilisateur l'effectif N des lots, l'effectif n des échantillons et le nombre d'acceptation c du plan de base ($c< =5$). Il demande ensuite à l'utilisateur d'entrer le taux d'erreurs d'entrée p hypothétique. Il calcule Pr(accept.), l'AOQ, l'AOQL (et le p correspondant à l'AOQL) et l'EIR pour tous les plans d'échantillonnage dans lesquels les effectifs des échantillons sont dans l'intervalle (n-3) à (n+5) et les nombres d'acceptation sont dans l'intervalle (c-1) à (c+2).

3.4. GRASP : Programme général d'évaluation de plans d'échantillonnage par attributs

Ce programme, publié par Schilling, Sheesley et Nelson [10], calcule la courbe d'efficacité d'un plan d'échantillonnage arbitraire simple, double ou multiple en utilisant la distribution hypergéométrique, binomiale, de Poisson ou normale. Il calcule également les taux d'éléments défectueux (percentiles) associés à des probabilités d'acceptation arbitraires spécifiées.

Pour l'utiliser, il faut être relié à une imprimante pour y imprimer les résultats. Le programme demande d'abord l'effectif N des lots. On peut entrer un effectif fini ou infini. Ensuite, il demande d'entrer le type de plan d'échantillonnage (échantillonnage simple, double, multiple), le nombre de taux d'éléments défectueux p pour lesquels le plan doit être évalué, le nombre de probabilités d'acceptation arbitraires à calculer, la distribution de probabilités à utiliser (hypergéométrique, binomiale, de Poisson ou normale).

L'utilisateur peut aussi laisser le programme choisir la distribution appropriée selon les plans à évaluer.

Ensuite, il faut entrer à chaque phase l'effectif des échantillons n, le nombre d'acceptation c, le nombre de rejet r (r=c+1 pour un plan simple), le taux d'éléments défectueux pour lequel le plan d'échantillonnage doit être évalué, l'erreur admissible sur la probabilité d'acceptation et la probabilité d'acceptation arbitraire choisie.

Un plan sommaire se composant de l'effectif des échantillons, du nombre d'acceptation et du nombre de rejet est imprimé à chaque phase. Des tables donnent la probabilité d'acceptation, la probabilité de rejet, et la probabilité d'acceptation ou de rejet des lots à chaque phase du plan d'échantillonnage. Enfin, un sommaire des caractéristiques d'efficacité du plan s'imprime. Ce sommaire fournit le taux d'éléments défectueux p, la probabilité d'acceptation des lots P(accept.), le nombre moyen d'unités vérifiées par échantillon ASN, la qualité moyenne après contrôle AOQ et le nombre moyen d'unités vérifiées par lot ATI.

3.5. PAIRIS : Programme d'analyse de l'impact des stratégies de contrôle avec rectification

Le programme [9] permet d'analyser l'impact d'une stratégie de re-contrôle de la qualité (Re-CQ) spécifiée. Rappelons que cette technique consiste à procéder à une seconde vérification par échantillonnage, plus sévére, d'un lot qui n'a pas été accepté lors de la première vérification. Entre les deux vérifications, le lot rejeté est renvoyé au préposé qui l'a traité pour qu'il corrige ses erreurs. Le programme s'applique à une politique de rectification utilisée avec un plan d'échantillonnage de contrôle initial et un plan d'échantillonnage Re-CQ. Le programme suppose qu'un seul contrôle Re-CQ est effectué sur chaque lot rejeté.

Le programme demande premièrement à l'utilisateur l'effectif N des lots, le plan d'échantillonnage du contrôle initial (n1 et c1) et le taux d'erreurs d'entrée hypothétique p. Ensuite, il demande trois taux de rectification s'appliquant aux lots soumis au contrôle initial, soit:

- Le degré de rectification des échantillons des lots acceptés par le contrôle initial;
- Le degré de rectification des échantillons des lots rejetés par le contrôle initial;
- Le degré de rectification des restes des lots rejetés par le contrôle initial.

Le programme demande alors à l'utilisateur de spécifier le plan d'échantillonnage Re-CQ (n2 et c2). Enfin, il demande trois taux de rectification supplémentaires s'appliquant aux lots soumis au contrôle Re-CQ, soit:

- Le degré de rectification des échantillons des lots acceptés par le contrôle Re-CQ;
- Le degré de rectification des échantillons des lots rejetés par le contrôle Re-CQ;
- Le degré de rectification des restes des lots rejetés par le contrôle Re-CQ.

Le taux d'erreurs d'entrée du contrôle Re-CQ est fonction de la rectification des restes des lots rejetés par le contrôle initial et du taux d'erreurs d'entrée du contrôle de la qualité. A l'aide de ces trois données, le programme calcule les caractéristiques suivantes à chaque phase du contrôle des échantillons: la probabilité d'acceptation Pr(accept.), la qualité moyenne après contrôle AOQ, le taux de contrôle probable EIR, et la limite de la qualité moyenne après contrôle AOQL.

Chacune des valeurs ci-dessus est calculée et affichée deux fois; premièrement pour la situation dans laquelle seul le plan d'échantillonnage du contrôle initial est utilisé et deuxièmement, pour celle où le plan Re-CQ l'est en plus du plan initial.

Le programme calcule et affiche la probabilité générale d'acceptation des lots résultant de l'application combinée du plan initial et du plan Re-CQ.
Enfin, la sortie comporte une section indiquant l'AOQ, l'AOQL et l'EIR du plan d'échantillonnage initial comme si celui-ci était un plan d'échantillonnage simple avec rectification complète.

3.6. Echantillonnage passe-lots: Calculateur de paramètres intermédiaires [11]

Ce programme a pour but de calculer en fonction des paramètres spécifiés une table de valeurs de Y utilisées comme paramètres intermédiaires dans les formules de l'échantillonnage passe-lots. Ces valeurs servent à déterminer les plans d'échantillonnage passe-lots et à calculer l'AOQL correspondant. Il sauvegarde cette table dans un fichier temporaire et peut l'imprimer.

Le module demande à l'utilisateur le plus petit et le plus grand intervalle de validation, le plus petit et le plus grand taux d'échantillonnage des lots, le plus petit et le plus grand taux d'échantillonnage à l'intérieur des lots n/N du plan de référence, l'incrément de ce taux et enfin, le plus petit et le plus grand nombre d'acceptation.

Il produit en fonction de ces données une table des valeurs de Y pour des niveaux donnés d'intervalles de validation et de taux d'échantillonnage des lots (fréquences) en fonction du taux d'échantillonnage à l'intérieur des lots et du nombre d'acceptation du plan d'échantillonnage de référence.

3.7. Autres programmes de calcul de plans d'échantillonnage

Pour des raisons techniques, certains programmes de plans d'échantillonnage ne peuvent pas être intégrés en tant qu'options en ligne du système. Ces modules doivent être exécutés séparément (souvent à cause de leurs besoins trop grands en mémoire). Voici une brève description de chacun permettant d'en obtenir les caractéristiques principales.

- Programmes de plans types

  Ce module est constitué de 3 programmes SAS servant à créer les tables de l'article «Standardization of QC Sampling Plans for Survey Operations» [1]. Ils devraient donc être utilisés pour améliorer les tables fournies dans cet article.

- Fonctions de plans d'échantillonnage

  Ces fonctions APL ont été initialement créées pour vérifier les plans d'échantillonnage types qui viennent d'être élaborés, mais on peut également s'en servir pour obtenir d'autres plans.

- Calculateur d'intervalles d'échantillonnage

  Ce module est utilisé pour l'opération de codage de la classification des activités économiques. Il calcule les taux d'échantillonnage et les points de changement des plans d'échantillonnage dans lesquels le prélèvement s'effectue de façon systématique sur des lots dont l'effectif est variable.

Pour tous les modules de ce logiciel, on peut obtenir plus de renseignements en consultant le document intitulé «Logiciel de plans d'échantillonnage de contrôle de la qualité (PLANSCQ), Aperçu» [12] qui décrit plus longuement la plupart d'entre eux.

4. Ajouts envisagés pour l'avenir

Le logiciel d'aide à la sélection des plans d'échantillonnage a été conçu de façon à ce qu'il soit relativement facile d'y ajouter de nouveaux modules à mesure que les besoins s'en font sentir. On pourra dans l'avenir développer un (ou plusieurs) nouveau module (ou l'obtenir d'une source extérieure) adapté à de nouvelles méthodes éventuelles d'échantillonnage pour certaines applications spécifiques et ensuite l'intégrer au système interactif en ajoutant les écrans d'assistance appropriés pour faciliter son utilisation.

Pour l'instant, une deuxième phase à ce projet est déjà commencée. Il s'agit de la réalisation d'un gestionnaire de fichiers des plans d'échantillonnage déjà existants (plans types et plans spécifiques utilisés pour des opérations particulières). Ceci permettra de regrouper l'ensemble des plans actuellement disponibles dans la SMAQ et de faciliter leur accès par les méthodologistes et le personnel de soutien des opérations.

Conclusion

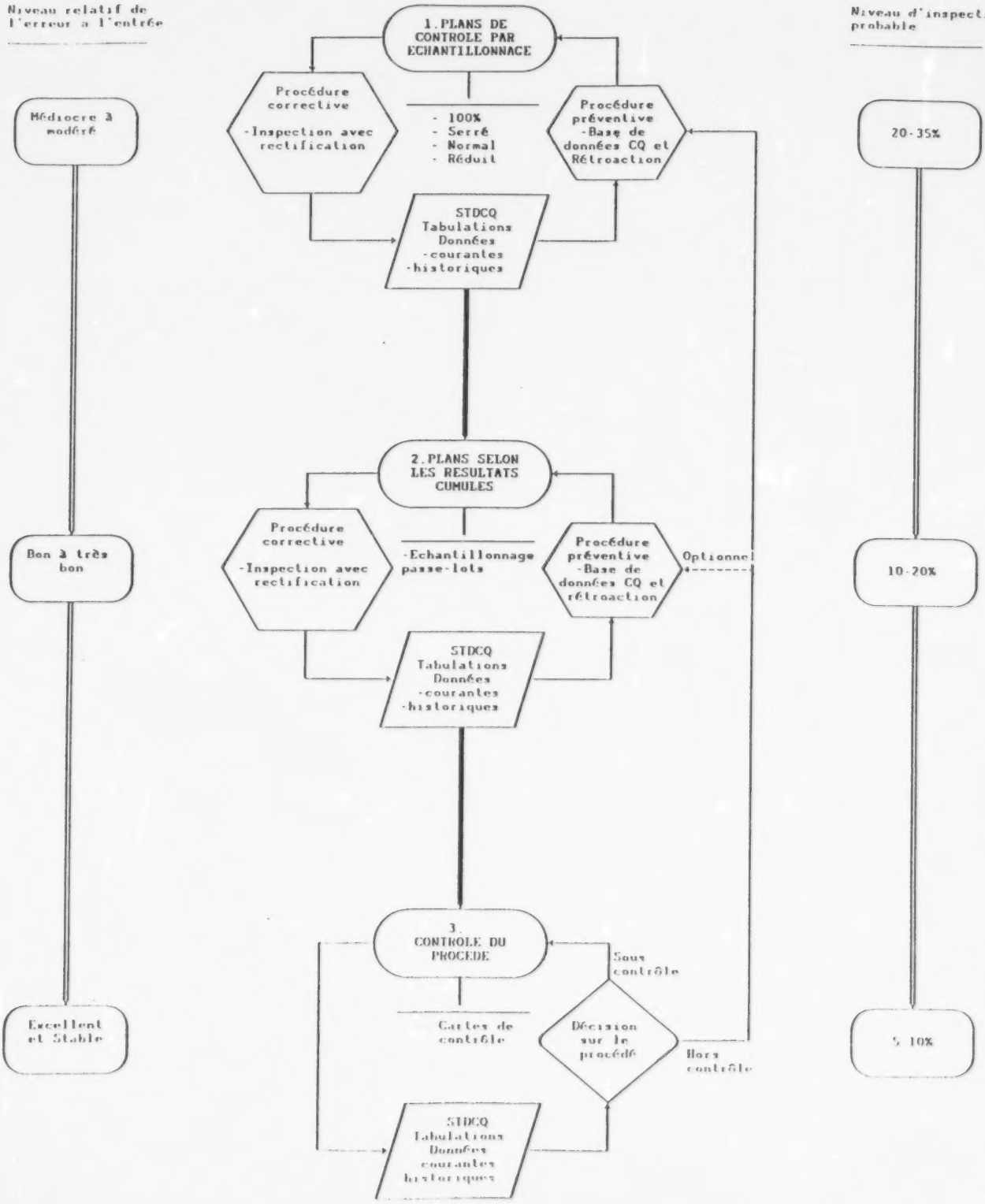
Ce logiciel de plans d'échantillonnage qui automatise les calculs reliés à la sélection des plans devrait aider à simplifier le travail du méthodologiste dans l'avenir. Des logiciels de ce genre s'avèrent habituellement être des outils importants pour l'amélioration des méthodes de travail ainsi que de leur efficacité. Il apparaît donc essentiel que le développement des systèmes informatiques et généralisés se poursuive pour une meilleure gestion des ressources disponibles et un maintien de la qualité des résultats et des produits statistiques offerts à la population.

Remerciements

L'auteur aimerait remercier sincèrement Walter Mudryk et Normand Laniel pour la révision technique de cet article ainsi que sa mère Nadia et son amie France pour la révision de l'orthographe. Tous les commentaires ou questions peuvent être adressés à l'auteur. Un effort sera fait pour y donner suite.

Annexe 1

Schéma de la stratégie globale de contrôle statistique de la qualité



NOTE: STDCQ - Système de traitement des données pour le contrôle de la qualité

BIBLIOGRAPHIE

1. Smith, J., Mudryk, W., Stankewich, R. - Standardization of QC Sampling Plans for Survey Operations; Part I: Guidelines and Rationals & Part II: Tables - Document interne, Statistique Canada, Ottawa, 1989.

2. Direction de la méthodologie - Cours de méthodologie d'enquêtes (416F); Manuel de l'étudiant - Document interne, Statistique Canada, Ottawa, mars 1990, version provisoire.

3. Statistique Canada - Lignes directrices concernant la qualité - Ottawa, Canada, avril 1987, deuxième édition.

4. Schilling, E.G. - Acceptance Sampling in Quality Control - Marcel Dekker, New York, 1982.

5. Dodge, H.F., Romig, H.G. - Sampling Inspection Tables; Single and Double Sampling - John Wiley & Sons, New York, 1959, deuxième édition.

6. MacMillan, J.H., Mudryk, W.V. - A non-parametric empirical Bayes Approach for Estimating a Process Average in Quality Control - Cahier de travail numéro BSMD-88-027E, DMEE, Direction de la méthodologie, Statistique Canada, Ottawa, août 1988.

7. Mudryk, W.V. - Système de gestion de la qualité dans les enquêtes - Techniques d'enquêtes, Vol. 14, no. 2, pp. 327-334, Statistique Canada, Ottawa, décembre 1988.

8. Stephens, K.S. - How to Perform Skip-Lot and Chain Sampling - Volume 4, ASQC Basic References in Quality Control: Statistical Techniques, American Society for Quality Control, Milwaukee, Wisconsin, 1982.

9. Smith, J. - L'auteur de la première version de ces programmes.

10. Schilling, E.G., Sheesley, J.H., Nelson, P.R. - GRASP: A General Routine for Attribute Sampling Plan Evaluation - Journal of Quality Technology, volume 10, numéro 3, juillet 1978.

11. Perry, R.L. - A System of Skip-Lot Sampling Plans for Lot Inspection - Thèse de doctorat, Rutgers University, New Brunswick, New Jersey, octobre 1970.

12. Labillois, T. - Logiciel de plans d'échantillonnage de contrôle de la qualité (Plans CQ); Aperçu - Document interne, Statistique Canada, Ottawa, avril 1990.

RECONSTITUTION DES MENAGES QUEBECOIS EN 1986 A L'AIDE DES DECLARATIONS DE REVENUS DE 1986

Nathalie Loignon, Robert Paré, Sylvain Veillette[1]

Introduction

Le Ministère du Revenu du Québec a entrepris au début de 1989 un projet visant à reconstituer les ménages québécois de 1986 afin d'évaluer l'impact des politiques fiscales touchant les ménages et d'étudier le comportement fiscal des membres d'un ménage.

La méthodologie utilisée pour reconstituer ces ménages est basée sur les données des déclarations de revenus de 1986. La description de cette méthodologie fait l'objet de la première partie de ce rapport alors que la seconde partie présente les résultats obtenus pour la C.U.Q. en plus de faire certaines comparaisons avec des données du recensement de 1986.

1. DESCRIPTION DE LA METHODOLOGIE

1.1 Les ménages d'après le fichier des déclarations de revenus

La définition de ménage retenue diffère légèrement des définitions habituellement utilisées :

> Epoux et épouse (avec ou sans enfants jamais mariés, de moins de 21 ans ou de 21 ans et plus s'ils sont déficients ou aux études postsecondaires) ou parent seul (peu importe son état matrimonial) avec un ou plusieurs enfants jamais mariés (de moins de 21 ans ou de 21 ans et plus s'ils sont déficients ou aux études postsecondaires) vivant dans le même logement.

La méthode utilisée pour reconstituer l'ensemble de ces ménages québécois à l'aide des informations contenues dans les 4 millions de déclarations de revenus s'effectue en deux étapes principales :

1. La normalisation des adresses par le code postal
2. L'appariement des ménages du Québec

Chacune de ces étapes comporte plusieurs traitements essentiels que nous présentons en détails.

---

[1]. Direction des études, recherches et statistiques, Ministère du Revenu du Québec.

1.2 La normalisation des adresses du fichier des déclarations de revenus

Cette étape est nécessaire puisque la méthodologie d'appariement repose en grande partie sur le fait que pour constituer un ménage les personnes doivent habiter à la même adresse. Afin de normaliser les adresses du fichier des déclarations de revenus à un niveau suffisant pour qu'une comparaison par ordinateur soit possible lors de l'appariement des ménages, deux traitements sur ces adresses sont essentiels: la structuration et la normalisation des adresses par le code postal.

Le premier traitement consiste à décomposer les deux lignes d'adresse des déclarations de revenus en champs d'information structurés. Le second consiste à déterminer à partir du code postal de l'adresse et du fichier des codes postaux du Québec (version corrigée et augmentée du fichier de Poste Canada), la municipalité et le nom de la rue (régions urbaines seulement) d'une adresse donnée. Voici deux exemples illustrant l'étape de normalisation des adresses :

Adresse du fichier des déclarations de revenus		Adresse structurée et normalisée			
RÉGION RURALE					
1) adr1 = 218 Marg Bourgeois CP 352		no. civique = 218		rue	= Marg Bourgeois
adr2 = Uneville Qué		no. app. =		casier postal	= 352
code postal = A1A 2B2		code postal = A1A 2B2		municipalité	= Uneville
RÉGION URBAINE					
6) adr1 = 818 Marg Bourgeois #1		no. civique = 818		rue	= Marguerite Bourgeois
adr2 = Céleri Tibec		no. app. = 1		casier postal	=
code postal = C1C 2D2		code postal = C1C 2D2		municipalité	= Tibec

La normalisation des adresses, bien que généralement très efficace pour les discriminer, s'avère inutile dans les cas suivants:

- adresses incomplètes: par exemple, lorsque le numéro d'appartement ou le casier postal n'est pas inscrit; lorsque le numéro civique comporte une lettre (si 118A est le numéro civique, seul le numéro 118 apparaît à l'adresse normalisée); enfin, lorsque le code postal est absent;

- adresse en région rurale: lorsque les noms de rue sont orthographiés différemment par des membres d'un même ménage, ils ne sont pas considérés comme demeurant à la même adresse; nous examinerons la possibilité de classer comme identiques des adresses normalisées qui sont partiellement en accord.

1.3 L'appariement des ménages du Québec

Cette fonction vise à reconstituer l'ensemble des ménages du Québec et obtenir, à partir du fichier des déclarations de revenus avec adresses normalisées, une banque d'information sur l'ensemble des ménages québécois et sur les personnes qui les composent. La reconstitution des ménages est basée sur des règles d'appariement regroupées en sept étapes successives. Les règles d'appariement sont appliquées sur un nombre de déclarations qui diminue d'une étape à l'autre puisque seules les déclarations non classées à une étape sont admissibles à l'étape suivante de la chaîne de traitement.

Une telle approche a été adoptée pour augmenter les chances de réussite de l'appariement car la probabilité d'une erreur dans l'appariement est plus grande à la dernière étape qu'à la première.

Principales étapes de l'appariement

(1) Premier appariement de conjoints : Couples mariés ou couples vivant en union libre

Le premier appariement de couples (mariés ou vivant en union libre) dont les deux conjoints produisent une déclaration de revenus est basé uniquement sur la présence des NAS (numéro d'assurance sociale) du déclarant et du conjoint dans la partie identification de la déclaration. Ainsi, les couples dont les deux contribuables déclarent mutuellement leur NAS et celui de leur conjoint ou les couples pour lesquels un seul NAS de conjoint est absent sont automatiquement classés comme couples mariés si leur état civil est marié ou comme couples vivant en union libre autrement.

La déclaration du NAS du conjoint n'étant pas obligatoire, certains couples dont les deux membres font une déclaration de revenus ne peuvent être appariés par NAS. Deux autres étapes sont ainsi nécessaires pour tenter d'apparier les couples mariés et les couples vivant en union libre qui restent . L'adresse constitue la principale clé de ces autres appariements.

(2) Deuxième appariement de conjoints : Couples mariés

Le deuxième appariement concerne uniquement les couples dont les deux conjoints se déclarent mariés. C'est ainsi que deux personnes mariées, de sexe différent, ayant obligatoirement été classées à la même adresse après l'étape de normalisation sont considérées comme un couple marié.

(3) Troisième appariement de conjoints : Couples vivant en union libre

Pour que deux personnes soient considérées comme un couple vivant en union libre l'ensemble des conditions suivantes doivent être satisfaites :

- Habiter à la même adresse
- Etre de sexe différent
- Avoir des noms de famille différents
- Avoir une différence d'âge entre les deux membres du couple inférieure ou égale à 15 ans.

Par contre, si une personne peut faire partie de plus d'un couple, aucun couple n'est formé.

Les étapes 1 à 3 ne peuvent apparier que les couples dont les deux conjoints font une déclaration de revenus.

(4) Identification des familles monoparentales

La présence d'une exemption pour famille monoparentale ou d'une exemption de soutien de famille dans la déclaration d'une personne non appariée jusqu'à cette étape sert à identifier une famille monoparentale.

(5) Identification des couples (mariés ou vivant en union libre)

Cette étape consiste à former des couples (mariés ou vivant en union libre) lorsqu'un seul conjoint remplit une déclaration de revenus. Si l'état civil de la personne déclarante est marié, on vérifie si une des informations suivantes est présente dans sa déclaration pour former un couple marié :

- Le NAS du conjoint
- Le revenu du conjoint
- Une exemption pour personne mariée
- Un transfert de déductions entre conjoint

On peut également utiliser l'information concernant l'identification du conjoint si le déclarant demande un remboursement d'impôts fonciers ou un crédit d'impôt pour taxes à la consommation pour identifier des couples mariés (si l'état civil du déclarant est marié) ou des couples vivant en union libre lorqu'un seul conjoint remplit une déclaration de revenus.

(6) Détermination des personnes seules

La plupart des contribuables de ce groupe devraient être des personnes seules. Cependant, il peut y avoir des contribuables qui devraient

normalement faire partie d'un des groupes précédents mais qui n'ont pu être classés compte tenu des informations disponibles sur leur déclaration. C'est pourquoi nous effectuons malgré tout l'étape 7 sur ces " personnes seules " qui consiste à déterminer la présence d'enfants à charge. Dans le cas échéant, ces " personnes seules " sont alors reclassées dans les familles monoparentales.

(7) **Détermination des enfants à charge** (déclarant ou non déclarant)

Cette étape, qui consiste à relier les enfants à leurs parents respectifs est réalisée en deux parties à chaque fois qu'un ménage est créé. La première partie permet de déterminer les enfants à charge qui remplissent une déclaration de revenus. Pour qu'une personne déclarante soit considérée comme un enfant elle doit d'abord être célibataire, avoir plus de 15 ans de différence avec sa mère et bien évidemment habiter à la même adresse que ses parents. De plus, si le ménage pour lequel on cherche à déterminer les enfants est un couple marié ou un couple vivant en union libre, il doit y avoir concordance entre les noms de famille (si les noms de famille sont composés, au moins un des noms de famille de l'enfant doit correspondre à au moins un des noms de famille d'un des parents). L'âge maximum d'un enfant est 21 ans à moins qu'un de ses parents demande une exemption pour personnes à charge aux études postsecondaires à plein temps ou une exemption pour personnes à charge agées de 21 ans ou plus et atteintes d'une infirmité physique ou mentale.

La deuxième partie consiste à déterminer les enfants à charge non déclarants en utilisant cette fois uniquement l'information sur les exemptions personnelles contenues dans la déclaration des parents. Encore une fois, l'âge limite d'un enfant non déclarant est assujetti aux mêmes conditions que celui d'un enfant déclarant. De plus, si la personne est âgée de 21 ans ou moins, on vérifie si les parents demandent l'exemption pour autres personnes à charge, l'exemption de soutien de famille, l'exemption pour famille monoparentale ou la déduction relative aux personnes atteintes d'une déficience physique ou mentale pour établir la présence d'enfants. D'autres éléments comme les allocations familiales provinciales ou fédérales, l'allocation pour enfants de moins de 6 ans et les frais de garde d'enfants peuvent également servir à déterminer les enfants à charge non déclarant.

2. RESULTATS OBTENUS DE LA RECONSTITUTION DES MENAGES DE LA C.U.Q.

L'année 1986 a été choisie pour développer une méthodologie de reconstitution des ménages du Québec parce que les données du recensement de 1986 sont disponibles et constituent donc la base de comparaison la plus adéquate. Cependant, il faut noter que le recensement s'est fait en juin 1986 alors que les informations des déclarations de revenus de 1986 sur la composition du ménage couvrent une période qui peut s'étendre de janvier 1986 à avril 1987.

Les résultats sont présentés en trois sections:

2.1 Caractéristiques des ménages privés de la C.U.Q..
2.2 Caractéristiques de la famille de recensement dans les ménages privés de la C.U.Q..
2.3 Répartition de la population de la C.U.Q. selon le sexe et l'âge.

2.1 Caractéristiques des ménages privés de la C.U.Q.

Ces caractéristiques apparaissent aux tableaux 1 et 2. Le tableau 1 montre des résultats très près du recensement en ce qui concerne la population totale, le nombre de personnes et la constitution des ménages privés (logement privé occupé). Ces résultats ont été obtenus des 291 756 déclarations de revenus de la C.U.Q.. Le nombre de ménages privés avec 2 familles ou plus est beaucoup plus élevé que le recensement et cela est dû aux adresses différentes mais non distinguables par la normalisation des adresses.

Le tableau 2 illustre le même problème avec les ménages privés de 10 personnes et plus. Le nombre de ménages d'une seule personne est également surestimé probablement parce que des membres d'un ménage inscrivent une adresse différente sur leur déclaration (adresse du bureau par exemble au lieu du logement).

2.2 Caractéristiques de la famille de recensement dans les ménages privés

Le tableau 3 présente des résultats sur la composition des familles de recensement dans les ménages privés assez semblables au recensement. Le fait que nous ayons limité l'âge des enfants à moins de 21 ans s'ils ne font pas l'objet d'une déduction fiscale explique sans doute la sous-représentation des familles de 5 personnes et plus. Le tableau 4 démontre la très bonne couverture des familles époux-épouse et la sous-représentation des familles monoparentales. Celle-ci est probablement due au fait que les chefs de ces familles ont des revenus plus faibles et qu'ils sont moins susceptibles de produire une déclaration.

2.3 Répartition de la population de la C.U.Q. selon le sexe et l'âge

Les tableaux 5 à 8 présentent la répartition de la population de la C.U.Q. selon l'âge et le sexe.

On obtient par la reconstitution des ménages une très bonne représentativité des tranches d'âge entre 25 et 64 ans. C'est un constat auquel on s'attendait puisque les personnes de ces tranches d'âge constituent la très grande majorité des personnes avec des revenus. Les personnes âgées (65 ans et plus) sont sous-représentées ainsi que les personnes âgées de 20 à 24 ans (faibles revenus donc moins susceptibles de produire une déclaration).

On remarquera au niveau des données selon le sexe que cette information n'est pas disponible pour les enfants identifiés comme personne à charge. Le tableau 8 le démontre puisque la plupart des personnes avec sexe "absent" sont âgées de moins de 20 ans. C'est l'explication de la faible représentativité des personnes des deux sexes âgées de moins de 20 ans (tableaux 6 et 7).

TABLEAU 1: CARACTÉRISTIQUES DE LA C.U.Q

	A RECENSEMENT	B MRQ	C % ERREUR (A-B)/A
POPULATION TOTALE	463 360	432 532	7
PERSONNES DANS LES MÉNAGES PRIVÉS	449 790	429 713	4
MÉNAGES PRIVÉS	172 710	175 390	-2
MÉNAGE NON FAMILIAL	52 505	58 687	-12
1 FAMILLE DE RECENSEMENT	119 475	113 220	5
2 FAMILLES DE RECENSEMENT OU +	740	3 483	-370
TOTAL FAMILLES DE RECENSEMENT	120 945	120 451	0

DÉFINITION DU M.R.Q. : LOGEMENT PRIVÉ AVEC MOINS DE 12 PERSONNES.

TABLEAU 2: RÉPARTITION DES MÉNAGES PRIVÉS DE LA CUQ SELON LA TAILLE

	A RECENSEMENT	B MRQ	C % ERREUR (A-B)/A
NOMBRE TOTAL DE MÉNAGES PRIVÉS	172 710	175 388	-2
1 PERSONNE	42 005	55 983	-33
2 PERSONNES	51 910	47 513	8
3 PERSONNES	32 270	29 685	8
4-5 PERSONNES	42 510	37 443	12
6-9 PERSONNES	3 975	4 666	-17
10 PERSONNES ET PLUS	45	98	-118

DÉFINITION DU M.R.Q. : LOGEMENT PRIVÉ AVEC MOINS DE 12 PERSONNES

TABLEAU 3 CARACTÉRISTIQUES DE LA FAMILLE DE RECENSEMENT DANS LES MÉNAGES PRIVÉS CUQ

	A RECENSEMENT	B MRQ	C (A-B)/A
NOMBRE TOTAL DE FAMILLES DE RECENSEMENT DANS LES MÉNAGES PRIVÉS	120 945	114 737	5
TAILLE DE LA FAMILLE			
2 PERSONNES	48 035	48 521	-1
3 PERSONNES	30 195	28 034	7
4 PERSONNES	29 820	27 123	9
5 PERSONNES ET PLUS	12 890	11 059	14

DÉFINITION DU MRQ : LOGEMENT AVEC MOINS DE 12 PERSONNES

TABLEAU 4 CARACTÉRISTIQUES DE LA FAMILLE DE RECENSEMENT DANS LES MÉNAGES PRIVÉS DE LA CUQ SELON LA STRUCTURE ET LA PRÉSENCE D'ENFANTS

	A RECENSEMENT	B MRQ	C % ERREUR (A-B)/A
NOMBRE DE FAMILLES ÉPOUX-ÉPOUSE	101 480	99 696	2
SANS ENFANT	36 255	38 491	-6
AVEC ENFANT	65 225	61 204	6
1 ENFANT	24 625	24 205	2
2 ENFANTS	28 260	26 149	7
3 ENFANTS OU PLUS	12 330	10 850	12
NOMBRE DE FAMILLES MONOPARENTALES	19 470	15 041	23
1 ENFANT	11 790	10 030	15
2 ENFANTS	5 585	3 829	31
3 ENFANTS OU PLUS	2 120	1 182	44
ENFANTS DANS LES FAMILLES DE RECENSEMENT	151 410	133 387	12

TABLEAU 5 : RÉPARTITION DE LA POPULATION DE LA CUQ SELON L'AGE

	A RECENSEMENT	B MRQ	C %ERREUR (A-B)/A
POPULATION TOTALE	463 360	432 532	7
AGE PRESENT	463 360	419 169	10
0-4 ANS	25 685	21 842	15
5-9 ANS	29 530	25 728	13
10-14 ANS	29 950	25 691	14
15-19 ANS	33 310	29 159	12
20-24 ANS	44 705	39 844	11
25-34 ANS	86 755	82 665	5
35-44 ANS	73 150	70 083	4
45-54 ANS	50 575	47 267	7
55-64 ANS	43 295	39 354	9
65-74 ANS	29 230	26 090	11
75 ANS ET +	17 125	12 446	27

TABLEAU 6 : RÉPARTITION DE LA POPULATION DE LA CUQ SELON LE SEXE ET L'AGE

	A RECENSEMENT	B MRQ	C % ERREUR (A-B)/A
HOMMES, TOTAL	221 170	154 723	30
0-4 ANS	13 415	14	100
5-9 ANS	15 000	22	100
10-14 ANS	15 520	62	100
15-19 ANS	16 770	5 792	65
20-24 ANS	22 165	17 770	20
25-34 ANS	42 540	40 293	5
35-44 ANS	35 195	34 127	3
45-54 ANS	24 255	22 929	5
55-64 ANS	19 515	18 037	8
65-74 ANS	11 485	10 645	7
75 ANS ET +	5 275	4 308	18
ABSENT		724	

TABLEAU 7 : RÉPARTITION DE LA POPULATION DE LA CUQ SELON LE SEXE ET L'AGE

	A RECENSEMENT	B MRQ	C % ERREUR (A-B)/A
FEMMES, TOTAL	242 200	170 007	30
0-4 ANS	12 270	11	100
5-9 ANS	14 530	19	100
10-14 ANS	14 430	45	100
15-19 ANS	16 540	4 251	74
20-24 ANS	22 540	17 847	21
25-34 ANS	44 215	41 900	5
35-44 ANS	37 955	35 857	6
45-54 ANS	26 320	24 282	8
55-64 ANS	23 780	21 239	11
65-74 ANS	17 745	15 416	13
75 ANS ET +	11 850	8 056	32
ABSENT		1 084	

TABLEAU 8 : RÉPARTITION DE LA POPULATION DE LA CUQ SELON LE SEXE ET L'AGE

	MRQ
SEXE ABSENT, TOTAL	109 320
0-4 ANS	22 034
5-9 ANS	25 918
10-14 ANS	25 805
15-19 ANS	19 774
20-24 ANS	3 405
25-34 ANS	485
35-44 ANS	99
45-54 ANS	56
55-64 ANS	78
65-74 ANS	29
75 ANS ET +	82
ABSENT	11 555

La méthodologie de la composante aréolaire de l'Enquête nationale sur les fermes

Jenny Lynch[1]

Résumé

L'Enquête nationale sur les fermes comporte deux types de bases de sondage : soit liste et aréolaire. Pour cette dernière, un plan d'échantillonnage à deux degrés est utilisé, où les unités sélectionnées au second degré sont des segments de terrain. Dans cet article, nous discuterons des aspects pratiques reliés à l'échantillonnage aréolaire, nous décrirons de plus les procédures cartographiques impliquées, ainsi que les problèmes et difficultés rencontrés lors de la collecte des données et de l'estimation.

1. Introduction

L'Enquête nationale sur les fermes (ENF) est une enquête annuelle menée de mai à juillet car elle comporte plusieurs composantes. Elle est effectuée dans toutes les provinces, sauf Terre-Neuve où une enquête séparée a lieu.

Les objectifs de l'enquête sont de fournir des estimations fiables et actuelles au niveau provincial pour essentiellement trois catégories de variables:

- les cultures (les superficies ensemencées)
- le bétail (la taille des cheptels)
- la situation financière (revenus et dépenses d'exploitation).

La population cible inclut toutes les fermes dont la vente de produits agricoles est supérieure à 250 $ durant les douze derniers mois. La population enquêtée exclut les fermes d'institutions ainsi que les fermes dans les réserves indiennes.

2. Les bases de sondage

Dans l'ENF nous retrouvons deux types de base de sondage. Le premier type est une base de liste où les unités sont les fermes du recensement agricole de 1986. Le nombre de bases de liste varie selon la province. Dans les provinces de l'ouest, nous avons deux bases de liste: la première (L1) comprend les grandes et moyennes fermes alors que la deuxième (L2) comporte les petites fermes. Dans les autres provinces, une seule base de liste est employée et renferme toutes les fermes du recensement.

L'autre type de base est aréolaire et les unités qui la composent sont des aires géographiques, plus précisément ce sont des secteurs de dénombrement agricoles

---

1 Jenny Lynch est méthodologiste à Statistique Canada, Section des enquêtes agricoles, Division des méthodes d'enquêtes-entreprises, Ottawa, Ontario.

(SDA). Un SDA est une région géographique assignée à un agent recenseur. On le définit agricole si au moins une ferme y a été énumérée lors du recensement.

Dans la base aréolaire, nous excluons tout SDA faisant partie de régions agricoles marginales comme dans le grand nord ou à la périphérie des villes. La base aréolaire couvre le Québec, l'Ontario et les provinces de l'ouest.

Le rôle de la base aréolaire est d'assurer une couverture complète de la population enquêtée car elle peut tenir compte des fermes qui ont été oubliées lors du recensement (sous-dénombrement) ou encore des nouvelles exploitations. Ce qui n'est pas le cas de la base de liste. De plus, dans les provinces de l'Ouest la base aréolaire sert de complément à la base de liste L1 (grandes et moyennes fermes).

Originalement dans les provinces de l'ouest nous ne voulions qu'une base de liste étant donné que la base aréolaire sert de complément à L1. A cause de contraintes de temps pour les publications, il était impossible de recueillir les données de la base aréolaire à temps. Nous avons alors décidé de créer une deuxième base de liste pour couvrir les petites fermes.

3. Méthode d'échantillonnage aréolaire

La méthode ici décrite s'applique tout particulièrement aux provinces de l'Ouest où un remaniement de l'enquête a été effectué en 1988. Un survol seulement de la méthodologie sera présenté car ce sujet fut discuté en détail lors de l'ACFAS de 1989 (voir Julien et Maranda (1989)).

Le plan de sondage en est un d'échantillonnage stratifié à deux degrés, où au premier degré les unités sont les SDA alors que les unités du deuxième degré sont des segments à l'intérieur des SDA choisis.

Pour la stratification, nous agrégeons au niveau de chaque SDA, les données de variables-clés. Cependant, nous utilisons seulement les données des fermes n'appartenant pas à la base L1 afin de rendre la base aréolaire le plus efficace possible étant donné qu'elle sert de complément à L1.

Nous déterminons ensuite les SDA auto-représentatifs, c'est-à-dire, ceux qui contribuent de façon importante selon une ou des variables-clés à l'activité agricole du SDA. Ces SDA sont alors regroupés en une strate.

Les SDA restants sont répartis en plusieurs strates par classification multi-dimensionnelle selon des variables-clés qui ont été au préalable standardisées.

À cette étape, le nombre de strates n'est pas encore déterminé car nous utilisons un approche empirique en effectuant plusieurs stratifications et répartitions d'échantillon avec différents nombres de strates. Avec les coefficients de variation obtenus à chaque essai, nous tentons de trouver un compromis entre un coefficient de variation acceptable et le plus petit nombre de strates.

La répartition de l'échantillon se fait de façon optimale multivariée sur les mêmes variables-clés afin d'augmenter la précision au niveau provincial pour une taille d'échantillon donnée. Toutefois, une contrainte à la répartition de l'échantillon est d'avoir une taille qui soit un multiple de 4.

Nous voulons un multiple de 4 car lors de la sélection de l'échantillon de SDA, nous allons choisir 4 répliques indépendantes de façon circulaire systématique pour des fins d'estimation. En raison de l'indépendance de la sélection, nous pouvons retrouver un SDA donné dans plus d'une réplique. Dans ce cas on s'assure de choisir des segments différents au second degré.

Pour l'ensemble des provinces couvertes par la base aréolaire nous retrouvons les tailles suivantes :

	nbre SDA ds base	nbre de strates	nbre SDA auto-rep	nbre SDA échant.	nbre seg échant.
Total	8719	178	19	1271	1770

4. Aspect cartographique

Une fois les SDA choisis, intervient l'aspect cartographique. Nous commençons alors par vérifier sur des cartes appropriées les exclusions possibles (par exemple les parcs provinciaux).

Nous repérons ensuite sur des cartes ayant les limites des SDA, les SDA choisis. On retrace sur les cartes topographiques le contour des SDA. Nous divisons alors chaque SDA en segments d'environ 3 milles carrés (7.5 km carrés) en utilisant des frontières naturelles autant que possible. On numérote les segments et on sélectionne par échantillon aléatoire simple sans remise des segments dans la proportion de 1/30. Dans l'exemple de l'annexe A, pour le SDA 711-056, le segment 13 a été sélectionné

Sur une même carte topographique, nous traçons tous les segments choisis. On assemble ainsi toutes les cartes d'une province (environ 15) pour créer une "énorme carte" que l'on envoie dans les bureaux régionaux afin qu'ils assignent la tâche de chaque interviewer.

On refait pour chaque interviewer, une carte de la tâche qu'il aura à effectuer (voir annexe B). Habituellement la tâche de l'interviewer varie entre 5 et 6 segments. De plus, nous envoyons à chacun, une carte à plus grande échelle de chaque segment qu'il aura à énumérer. À l'annexe C, le segment 5 de l'annexe B est illustré. Une carte de délimitation est aussi incluse. En fait pour les provinces de l'ouest, c'est le plus souvent une grille divisée en sections, sinon c'est une photographie aérienne. Cette carte sert à délimiter le contour des

fermes à l'intérieur du segment (voir annexe D).

5. La collecte

La collecte se fait par une entrevue en personne. L'interviewer arrive dans le segment et doit obtenir l'information auprès de tous les fermiers qui ont leur ferme, en totalité ou en partie dans le segment. On retrouve en moyenne 5 fermes par segment. L'interviewer demande la superficie totale de la ferme et la superficie à l'intérieur du segment.

Par un appariement manuel sur place, l'interviewer détecte si le fermier appartient à la base de liste L1. S'il n'appartient pas à L1, on lui demande de remplir un questionnaire concernant les cultures, le bétail et les dépenses. Si le fermier appartient à L1, il n'est pas nécessaire de recueillir d'information car il est représenté par l'échantillon tiré de la base de liste. Les fermes de segments sont vérifiées à nouveau au bureau central en ce qui a trait au chevauchement des bases. Cette étape est très importante car lors de l'estimation nous voulons traiter les bases de liste et aréolaire de façon indépendante et mutuellement exclusive.

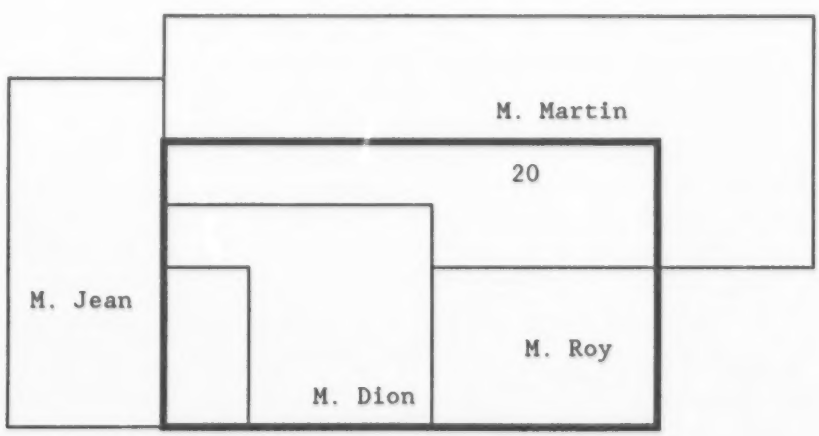
6. L'estimation

6.1 Totaux

Pour l'estimation des totaux nous devons déterminer la contribution de la ferme dans le segment soit la proportion W qui est spécifique à chaque ferme :

$$W_i = \frac{\text{superficie de la ferme dans le segment}}{\text{superficie totale de la ferme}}$$

Par exemple, si M. Martin a une ferme de 100 acres et que la partie de la ferme qui est à l'intérieur du segment est de 20 acres, alors toutes les données recueillies pour cette ferme seront multipliées par 20%.



On somme la contribution de toutes les fermes du segment afin d'obtenir le total pour le segment, soit :

$$\hat{Y} = \sum_k W_k y_k$$

Nous utilisons ensuite la méthode usuelle du plan à 2 degrés où M/m est le facteur d'expansion au niveau du segment et N/n est le facteur au niveau du SDA . Pour la réplique r nous avons :

$$\hat{Y}_h^{(r)} = \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} \hat{Y}_{hij}^{(r)}$$

On divise par 4 la somme des totaux des 4 répliques :

$$\hat{Y}_h = \frac{\sum_{r=1}^{4} \hat{Y}_h^{(r)}}{4}$$

6.2 Variance

Pour l'estimation de la variance, nous avons le problème suivant : il est fréquent de ne retrouver qu'un seul segment échantillonné par SDA. Il est alors impossible par les formules traditionnelles d'obtenir la variation au 2è degré. D'où l'utilisation des répliques afin de pouvoir calculer la variance. Nous utilisons donc la formule de la variance pour répliques qui est la suivante:

$$var(\hat{Y}_h) = \frac{1}{(4-1)4} \left(\sum_{r=1}^{4} (\hat{Y}_h^{(r)})^2 - \frac{\left(\sum_{r=1}^{4} \hat{Y}_h^{(r)} \right)^2}{4} \right)$$

Les estimations provinciales sont obtenues en additionnant les estimations des bases de liste et aréolaire.

7. Quelques problèmes

Un des problèmes rencontré dans la base aréolaire de l'enquête est la contribution de la ferme dans le segment. Il est en effet souvent difficile pour le fermier d'évaluer la portion de sa ferme dans le segment car les limites que nous créons ne lui sont pas toujours familières. Ceci peut entrainer une variation de nos estimations.

Un autre problème est le chevauchement entre la base liste et aréolaire qui parfois n'est pas détecté et contribue alors à du double compte.

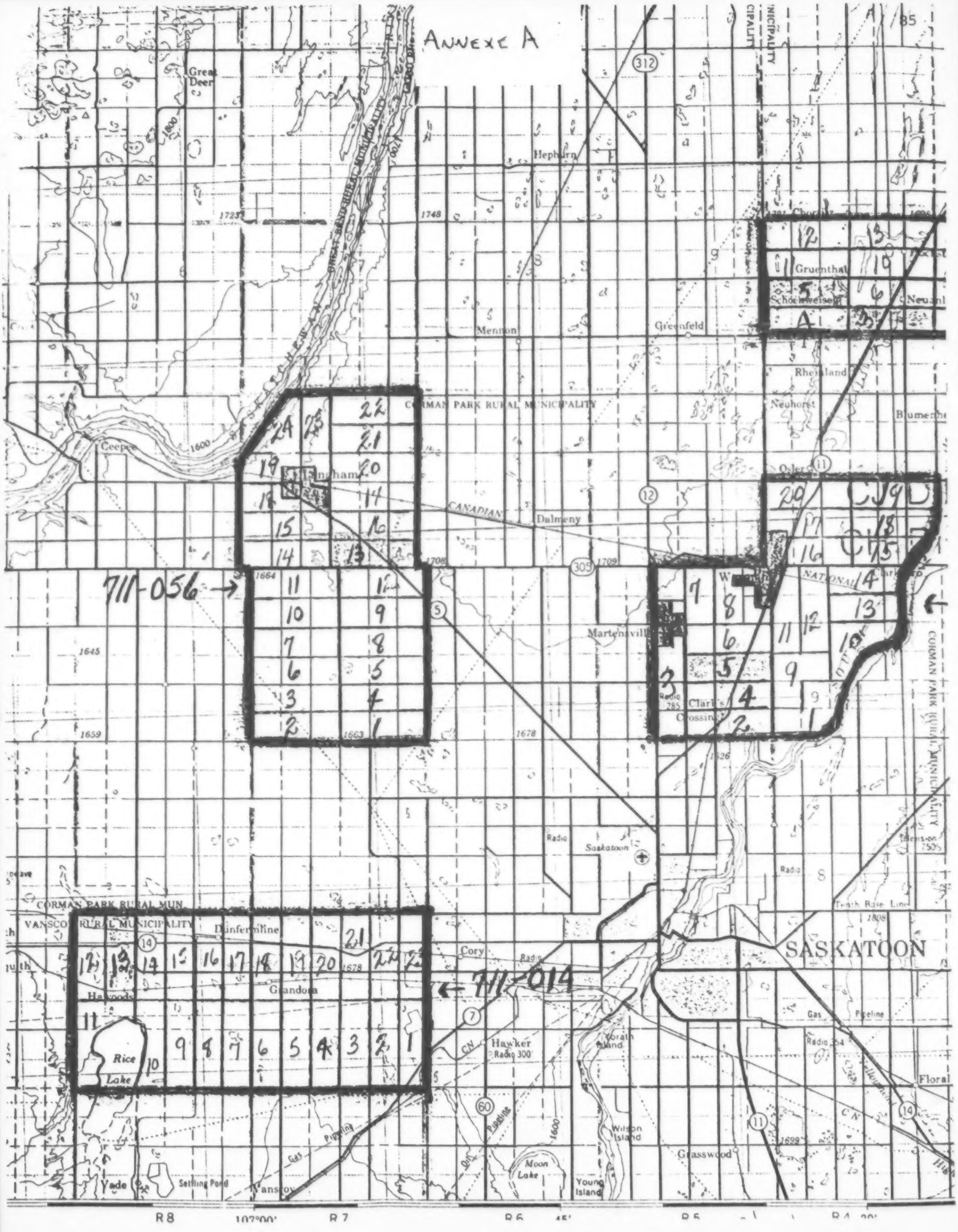
Un désavantage certain sont les coûts encourus. Pour la base aréolaire, il en coûte en moyenne 65$ pour un interview alors que dans la base liste où les interviews se font par téléphone, le coût est de 15$.

De plus, la base aréolaire est beaucoup moins efficace que la base liste. Elle contribue à seulement 20% de l'estimation totale provinciale et à 50% de la variation provinciale.

Mais, tant qu'il n'existera pas de mécanismes pour mettre à jour convenablement la base de liste, c'est le prix à payer pour assurer une couverture complète de la population enquêtée.

Référence

Julien, C., Maranda, F. (1989), "Le remaniement du plan de sondage de l'enquête nationale sur les fermes", ACFAS 1989.


Annexe A
Great Deer
Hepburn
Chortitz
Gruenthal
Schoenweise
Mennon
Greenfeld
Rheinland
Neuhorst
Osler
CORMAN PARK RURAL MUNICIPALITY
Langham
Dalmeny
CANADIAN
NATIONAL
711-056
Martensville
Clark's Crossing
Radio 285
Radio
Saskatoon
SASKATOON
CORMAN PARK RURAL MUN.
VANSCOY RURAL MUNICIPALITY
Dunfermline
Cory
Grandora
711-014
Rice Lake
Hawker
Radio 300'
Yorath Island
Wilson Island
Grasswood
Moon Lake
Young Island
Vanscoy
Settling Pond
Gas Pipeline
Tenth Base Line
Floral
R8
107°00'
R7
R6
45'
R5
R4
30'

Annexe B

Waldheim
Rabbit Lake
Great Deer
Hepburn
312
CORMAN PARK RURAL MUNICIPALITY
Langham
Dalmeny
Canadian
4
2
6
5
3
1
305
Warman
National
Clarkboro
Osler
11
Chortitz
Hochstad
Gruenthal
Schoenweiser
Neuanlage
Greenfeld
Rheinland
Neuhorst
Blumenheim
Mennon
Ceepee
CJ DJ
CH DH
Martensville
Clark's Crossing
Radio 285
ABERDEEN RURAL MUNICIPALITY
CORMAN PARK RURAL MUNICIPALITY
South
Saskatoon
Radio
Television 750
Tenth Base Line
CORMAN PARK RURAL MUN
VANSCOY RURAL MUNICIPALITY
Dunfermline
14
Cory
Earley
Haywoods
Grandora
Rice Lake
7
Hawker
Radio 300
60
Wilson Island
Grasswood
Flora
Gas Pipeline
SASKATOON
1664
1708
1709
1663
1678
1645
1659
1642
1748
1721
1626
1600
1699

Annexe C

Statistics Canada | Statistique Canada

1990

National Farm Survey | **Enquête nationale sur les exploitations agricoles**

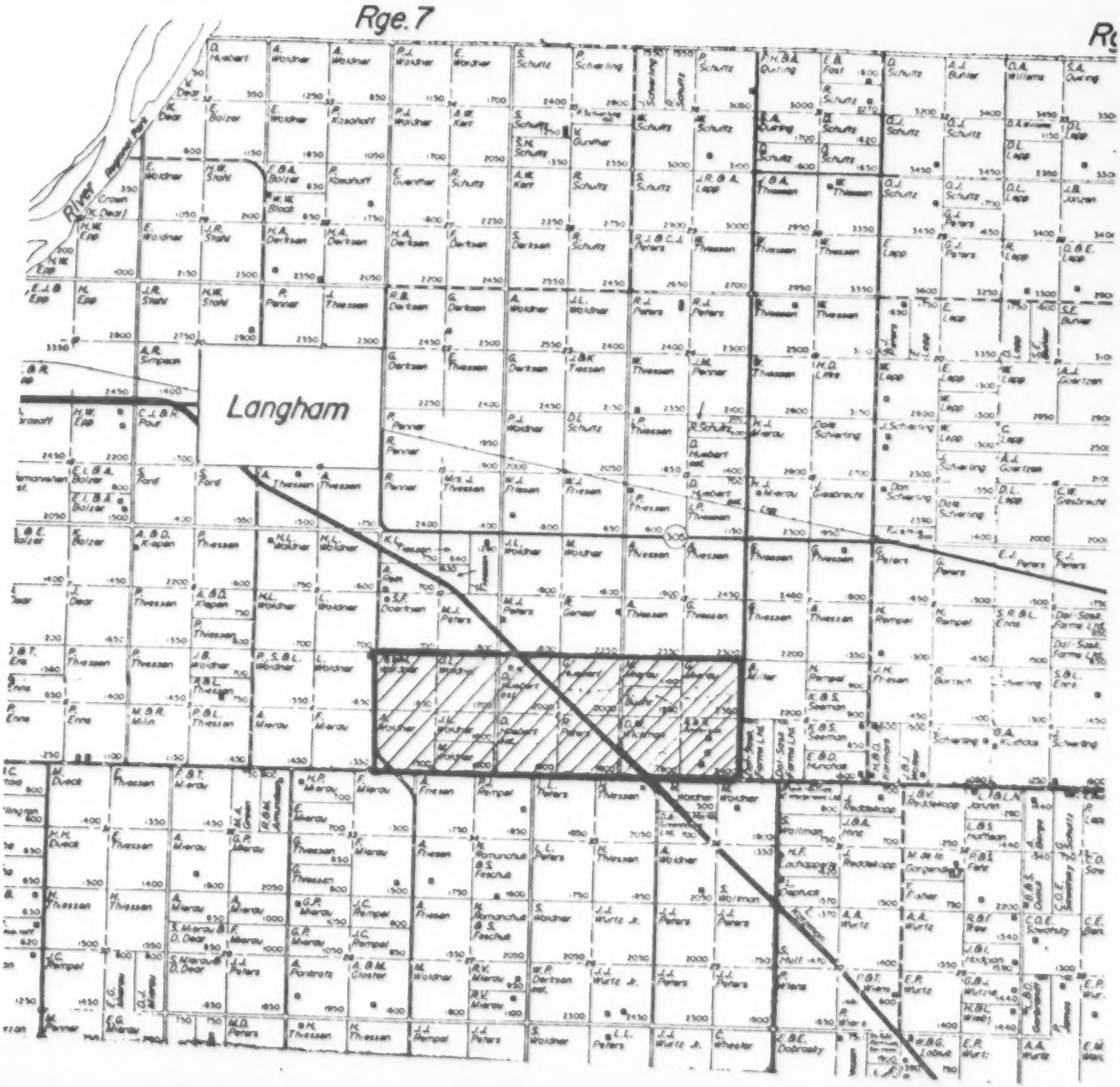
Segment Map | **Carte de segment**

PROV. 47

E.D. / C.E. 11

E.A. / S.D. 056

Segment No. / N° de segment 113



88

ANNEXE D

1990

Statistics Canada | Statistique Canada

National Farm Survey | **Enquête nationale sur les exploitations agricoles**

Plotting Map | **Carte de délimitation**

PROV. 47

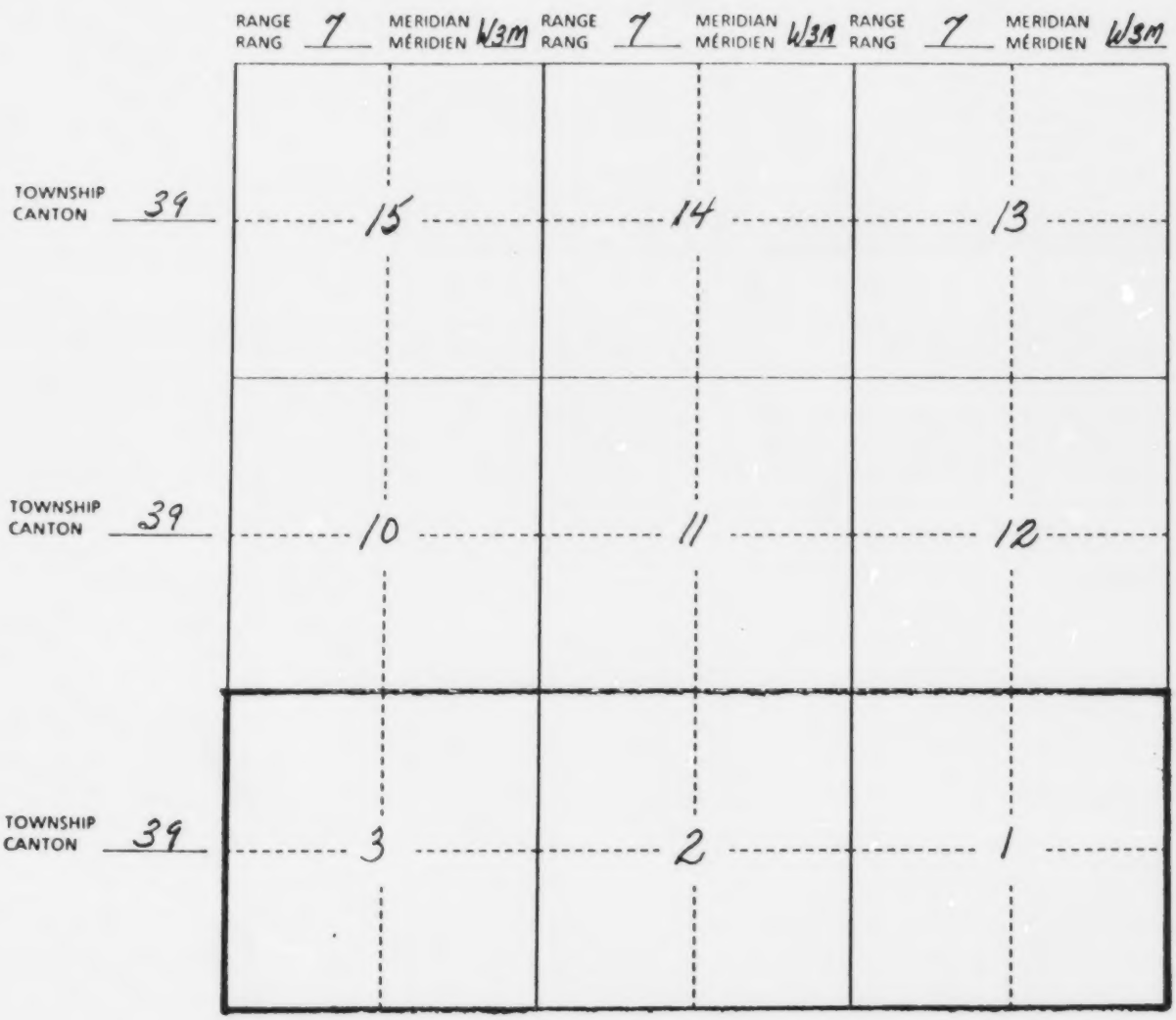
E.D. C.E. 11

E.A. S.D. 056

Segment No. N° de segment 113

Ce segment fait partie de l'échantillon cette année pour la TROISIEME fois.

This is the THIRD year in the sample for this segment.



LA MÉTHODOLOGIE DES ENQUÊTES SUR LE COMPORTEMENT ALIMENTAIRE DE GROUPES CIBLES DE LA POPULATION QUÉBÉCOISE

NICOLE A. SEOANE

Service des politiques alimentaires
et des relations commerciales
Ministère de l'Agriculture, des Pêcheries
et de l'Alimentation du Québec

La présente a pour but de fournir des informations au regard des enquêtes sur les comportements alimentaires menées au cours des dernières années par le ministère de l'Agriculture, des Pêcheries et de l'Alimentation (MAPA).

Ces enquêtes furent menées sous l'égide du Service des politiques alimentaires et des relations commerciales dont le mandat se résume ainsi:

> *élaborer des programmes destinés à développer ou améliorer sous toutes ses formes, la qualité de l'alimentation et des services alimentaires offerts à la population du Québec tout en y faisant la promotion des produits bio-alimentaires québécois.*

*À l'intérieur d'une stratégie globale, appelée " **Cadre général d'intervention en matière d'alimentation** ", des populations cibles ont été identifiées, leurs besoins et leurs préoccupations face à la saine alimentation cernés et une démarche précise délimitée pour des programmes et des activités du MAPA. L'approche par population cible permet d'adapter les interventions selon les besoins de chaque groupe, limite le champ d'action à des mesures portant davantage sur la modification des comportements alimentaires et est assez facile à opérationnaliser, particulièrement dans le contexte de ressources humaines limitées.*

À ce jour, trois populations ont fait l'objet d'interventions de la part du MAPA: les ménages québécois, les jeunes en milieu scolaire et les aînés.Le choix des populations cibles à privilégier pour nos programmes d'intervention s'est fait en consultant plusieurs sources d'information dont les résultats de recherche dans le domaine de la nutrition appliquée et communautaire, les enquêtes sur les tendances à la consommation, notamment celles menées par Agriculture Canada et les données démographiques. Les professionnels et partenaires du MAPA, les intervenants du milieu et parfois des groupes de consommateurs ont aussi participé à cette phase de consultation. De plus, avant d'arrêter une stratégie d'intervention précise, le MAPA a choisi de mener ses propres enquêtes afin d'actualiser le profil des habitudes alimentaires des segments de population où son engagement était jugé prioritaire.

Les trois enquêtes menées par le MAPA partageaient un objectif commun, soit:

de fournir des informations courantes sur les comportements alimentaires de populations cibles afin d'orienter ses activités de promotion de la saine alimentation.

Le tableau suivant indique sommairement quelques détails du plan général d'échantillonnage de chacune des trois enquêtes. Dans les trois cas, des partenaires externes ont collaboré à diverses étapes pour assurer la bonne marche des projets.

Tableau 1. Partenaires et populations visées

ENQUÊTES	PARTENAIRES	NOMBRE D'UNITÉS DE COLLABORATEURS	POPULATION VISÉE
ménages	les Cercles de Fermières du Québec	867	toutes les municipalités
jeunes	les Fédérations des Cercles de Fermières du Québec	26	114 écoles
aînés	le Conseil de l'Âge d'Or de la région de Québec	1	40000 membres

Dans un prochain temps, les enquêtes seront abordées une à une pour démontrer les différentes approches méthodologiques utilisées. Sur le plan des sciences de la statistique, ces approches comportent certes des failles, mais elles constituaient les meilleurs choix dans les circonstances. Le mode d'échantillonnage utilisé a été conditionné par le contexte des enquêtes et particulièrement par les partenaires du dossier.

LES MÉNAGES QUÉBÉCOIS

Pour les fins de l'étude, tout le Québec, sauf le territoire du Nouveau-Québec, a été divisé d'après la configuration géographique des 25 fédérations des Cercles de Fermières du Québec, ainsi que les trois comtés non représentés dans l'actuelle répartition de ces territoires, pour un total de 28 territoires.

INSTRUMENTS D'ÉCHANTILLONNAGE

- *Carte géographique*
- *Répertoire des municipalités du Québec 1981*
- *Grille de pondération*
- *Annuaire téléphonique*
- *Série de nombres au hasard*

À partir du " Répertoire des municipalités du Québec 1981 " , toutes les municipalités et leurs populations respectives se trouvant dans les limites de ces territoires ont été dénombrées. Le nombre de répondants à choisir a été déterminé à l'aide d'une grille de pondération. Chaque municipalité est donc représentée dans le plan d'échantillonnage proportionnellement à sa population.

Une série de nombres, choisis au hasard par informatique en tenant compte de la taille individuelle de chaque municipalité, a été fournie aux cercles locaux. Cette série a été appliquée aux annuaires téléphoniques locaux pour choisir les répondants. Dans le cas de la région Montréal, où le bottin est volumineux et ce système n'était pas très pratique pour nos collaborateurs volontaires, une section spécifique du bottin était assignée à chacun des 21 cercles de la région et un échantillonnage systématique fut effectué à l'intérieur de ces sections du bottin. Les détails du plan d'échantillonnage sont disponibles dans le rapport d'enquête. Il y a lieu de souligner que, lors d'un congrès de la Fédération des Cercles de Fermières du Québec, il y a eu présentation d'une session de formation comprenant des exercices pratiques pour familiariser nos intervenantes avec la méthodologie de l'enquête. De plus, une personne ressource était disponible tout au long de la période de collecte des données pour répondre aux questions et solutionner les problèmes qu'une enquête de cette envergure peut susciter.

Dans les 26 territoires d'échantillonnage, des membres désignées de cercles locaux ont pris contact avec les répondants. Chaque cercle téléphonait aux répondants de sa liste principale. D'après une formule type d'approche, il sollicitait leur participation. En cas de refus, les listes de répondants substituts étaient utilisées.

INSTRUMENTS DE MESURE

- *Questionnaire autogéré, environ 30 à 45 minutes*
- *Format cahier 8 1/2" X 11", 23 pages*
- *44 questions portant sur l'alimentation*
- *15 questions sociodémographiques*
- *7 questions portant sur le recyclage*

Un questionnaire constituait l'instrument de mesure le plus approprié à l'enquête. Il permettait d'atteindre facilement les 4 698 répondants

répartis sur le territoire québécois. Le questionnaire était livré personnellement ou par courrier dans les jours suivant l'appel. Quelques jours plus tard, les participant recevaient un appel téléphonique pour les encourager à répondre au questionnaire, si ce n'était déjà fait.

Le retour des questionnaires s'est effectué directement à Québec aux bureaux du MAPA par enveloppe-réponse prétimbrée. À partir des 4 698 expédiés, 3 285 furent retournés au Ministère dont 50 inutilisables parce qu'incomplets ou reçus trop tard. Ce sont donc 3 235 questionnaires qui ont constitué l'unité d'analyse donnant un taux de participation de 68,9%, et ce, malgré la longueur du questionnaire.

Tous les questionnaires furent analysés globalement afin d'obtenir un profil de l'ensemble des ménages participants. Par la suite, les répondants furent répartis selon le statut conféré à leur lieu de résidence, conformément au Répertoire des municipalités du Québec 1981. À cette fin, les municipalités dites de campagne et de village dans ce répertoire sont qualifiées "rurales" dans la présente enquête, alors que les municipalités de cité et de ville sont classifiées "urbaines". Cette répartition de l'échantillon fut motivée par les caractéristiques spécifiques de la population cible du Ministère. D'une part, le Ministère doit fournir aux agriculteurs, qui résident majoritairement dans les régions rurales, l'assistance technique, scientifique et économique nécessaire pour encourager une production agricole rentable. D'autre part, le Ministère doit fournir aux consommateurs québécois, regroupés principalement dans les centres urbains, les informations nécessaires pour les inciter à rechercher et à consommer davantage les produits bio-alimentaires québécois. Puisque les programmes du Service des politiques alimentaires et des relations commerciales s'adressent à toute la population, la répartition de l'échantillonnage devait permettre l'identification des caractéristiques propres du comportement alimentaire des ménages vivant soit en milieu rural, soit en milieu urbain.

L'ensemble de l'échantillon fut aussi divisé pour regrouper les répondants de deux grands centres urbains, soit les Communautés urbaines de Montréal et de Québec et ceux des autres régions du Québec. Étant donné le nombre important de grossistes et de détaillants en alimentation présents dans ces grands centres urbains, on a supposé que les consommateurs de ces régions pourraient avoir un comportement différent, hypothèse que nous avons voulu vérifier.

Des analyses de fréquences simples et croisées, le test du Khi deux pour les distributions de fréquence et le test "t" de Student furent effectués sur différentes parties des données.

ÉCHÉANCIER

Printemps 1983:	*Consultation avec les Cercles des Fermières*
	Élaboration du protocole
Août 1983:	*Atelier*
Octobre 1983:	*Mise à l'essai du questionnaire*
Décembre 1983 à mars 1984:	*Collecte des données*
Mai, juin 1984:	*Saisie, analyses préliminaires*
Juillet 1984:	*Interprétation et rédaction*
Août 1984:	*Lancement du rapport préliminaire*
Juin à octobre 1985	*Analyses détaillées*
Décembre 1985:	*Rapport final*

Les résultats furent dévoilés en deux temps. D'abord, il y a eu lancement d'un rapport préliminaire en août 1984, à peine 5 mois après la fin de la période de la collecte des données, ceci pour respecter un engagement auprès des Cercles de Fermières du Québec. En décembre 1985, le rapport complet comportant des analyses plus détaillées fut publié. Les résultats ont été largement diffusés dans le monde bio-alimentaire et de la santé, ainsi qu'auprès du public.

Certaines données ont servi d'assise à des programmes d'information aux consommateurs, à des activités de marketing et publicité ainsi qu'aux orientations préconisées pour le développement de nouveaux produits.

LES JEUNES EN MILIEU SCOLAIRE

Comme dans l'enquête précédente, tout le Québec, sauf le territoire du Nouveau-Québec, a été divisé d'après la configuration des territoires des 25 fédérations des Cercles de Fermières du Québec. Les trois comtés ne figurant pas dans cette répartition géographique ont été représentés par une personne membre de l'Exécutif provincial.

Le choix des établissements scolaires invités à participer à l'étude s'est fait au hasard à partir du "Répertoire des organismes et des écoles" du ministère de l'Éducation du Québec. Le nom des écoles choisies et une liste de substituts en cas de refus de particpation étaient fournis aux responsables du Comité Agriculture-Consommation de chaque fédération. Une moyenne de trois écoles secondaires et un CÉGEP était visée dans chaque fédération, mais une certaine flexibilité était de rigueur pour les régions à forte ou faible densité de population, notamment les communautés urbaines de Montréal, de Québec et les Iles-de-la-Madeleine.

La responsable du Comité Agriculture-Consommation dans chaque fédération assumait le suivi de l'enquête, selon un protocole prédéterminé. Munie de lettres explicatives, elle établissait le contact avec le directeur ou la directrice des écoles choisies sur son territoire. Pour minimiser la perturbation de l'horaire scolaire, on demandait à la direction de l'école de choisir au hasard une classe complète d'environ 30 à 40 élèves, ou deux classes plus petites dont l'inscription totalisait au moins 30 personnes. La classe constituait donc la véritable unité d'échantillonnage. La rencontre avec le groupe d'élèves participant se faisait selon les ententes déjà prises individuellement avec les autorités scolaires. Chaque élève recevait une lettre l'encourageant à participer et un livret de renseignements à remplir. La collecte des données s'est faite au moyen d'un livret de 24 pages spécialement conçu pour l'enquête et dont la disposition permettait le traitement informatisé des dossiers.

INSTRUMENTS DE MESURE

- *Livret questionnaire*
  - *Renseignements d'ordre personnel*
  - *Journal alimentaire (3 jours)*
  - *Habitudes de vie: tabac, alcool, sport, etc.*
  - *Connaissances*

Comme le laisse supposer l'objectif commun des enquêtes du MAPA énoncé précédemment, une partie importante du livret était prévue pour tenter d'identifier les habitudes alimentaires des jeunes. Les participants devaient donc inscrire tous les aliments et boissons consommés pendant trois jours. À l'intérieur de ce journal alimentaire, chaque journée était divisée en six périodes reflétant la répartition habituelle de la consommation alimentaire journalière. Des espaces étaient donc prévus pour noter des aliments et des boissons consommés au déjeuner, à la collation du matin, au dîner, à la collation de l'après-midi, au souper et à la collation du soir.

Compte tenu de la nature descriptive de l'étude, tous les aliments rapportés dans les journaux alimentaires ont été codés selon leur appartenance à 93 catégories d'aliments déterminées à l'avance. Ces catégories ont été élaborées pour permettre de déceler certaines habitudes de consommation (ex.: lait pris comme boisson ou lait dans les céréales; légumes crus ou légumes cuits, ...), en plus d'identifier les aliments de façon qualitative (ex.: lait 2% ou lait entier; pain blanc ou pain de blé entier; beurre ou margarine ...).

La nomenclature des catégories devait aussi permettre l'identification de certains aliments d'origine québécoise. Enfin, divers regroupements de ces catégories ont été faits pour donner une vue plus générale de la prise alimentaire et pour permettre la comparaison des résultats avec ceux d'autres études, notamment l'enquête Nutrition Canada.

ÉCHÉANCIER

Août 1983:	*atelier*
Oct. - déc. 1983:	*élaboration et validation du livret*
Février 1984:	*contact avec les écoles*
Mars-avril 1984:	*collecte de données*
Mai-août 1984:	*codification*
Septembre 1984:	*saisie*
Oct. - nov. 1984:	*analyses informatiques fréquences et test "t"*
Nov. 1984 - mars 1985:	*interprétation, rédaction*
Mai 1985:	*lancement, " Année internationale des Jeunes "*

Une conférence de presse ministérielle, tenue dans une école locale en compagnie de jeunes, a servi de tribune au lancement du rapport. Ce document a joui d'une large diffusion dans les milieux de la santé, de l'éducation, du monde bio-alimentaire en plus d'être l'objet d'une publication scientifique et de plusieurs communications dans les médias.

Les données ont servi d'assise à plusieurs actions du Ministère dont la plus importante, " L'alimentation au Québec: voies d'action pour le milieu scolaire", une orientation conjointe avec le ministère de la Santé et des Services sociaux et le ministère de l'Éducation du Québec. Le Québec est d'ailleurs la seule province qui dispose d'une orientation provinciale en matière de nutrition pour le milieu scolaire.

Pour sa part, le MAPA a élaboré une série d'activités pour appuyer cette orientation, dont les quatre volets se résument ainsi:

- *un programme d'activités en milieu scolaire primaire ayant pour thème la consommation de légumes,*
- *des guides d'activités au secondaire pour appuyer les cours ayant un contenu "alimentation",*
- *des services-conseils en matière d'achats pour les gestionnaires de services alimentaires scolaires,*
- *des services-conseils en matière d'hygiène et de salubrité pour les manipulateurs d'aliments des services alimentaires scolaires.*

Pour faire connaître ces quatre volets d'action du MAPA, des offres de services ont été expédiées aux commissions scolaires, leur expliquant le contenu, les objectifs ainsi que la procédure à suivre pour se prévaloir de ces services. La programme du MAPA a connu un succès tel que, en date du 1er janvier 1990, environ 38 000 bloc notes avaient été distribués aux élèves du primaire dans le cadre d'une Fête des légumes et un peu plus de 50 000 passeports ont été acheminés aux élèves du secondaire dans la cadre des cours d'économie familiale en secondaire deux.

LES AINÉS

L'échantillon a été choisi au hasard, parmi la liste des 40 000 membres du Conseil de l'Âge d'Or de la région de Québec. Le bureau du Conseil a participé à cette démarche en élaborant un programme informatisé permettant de faire une sélection aléatoire parmi ses membres et de délivrer des étiquettes préadressées en leur nom.

Par suite des commentaires recueillis lors de consultations auprès de groupes indépendants d'aînés et d'intervenants du milieu, un questionnaire autogéré a été privilégié comme instrument de mesure, malgré les faiblesses reconnues que cette méthode comporte. Cette méthode de cueillette d'information permettait aux participants de répondre de façon anonyme, selon leur rythme et leur horaire.

INTRUMENT DE MESURE

- *Questionnaire autogéré*
- *14 pages, cahier*
- *Caractère typographique gras*
- *Validation préalable*

Le questionnaire de quatorze pages, dans un caractère typographique approprié, en forme de cahier, présentait une cinquantaine de questions, majoritairement de type fermé. Les répondants n'avaient donc qu'à cocher la réponse appropriée, ce qui pouvait encourager une plus grande participation. Deux pages à la fin du questionnaire invitaient les répondants à partager tout commentaire jugé pertinent.

Le type d'informations recueillies a permis d'effectuer des analyses de nature qualitiative, dont les fréquences simples et croisées et le test du "khi deux". Tous les questionnaires furent analysés globalement pour obtenir un profil de l'ensemble des répondants. Par la suite, des sous-groupes ont été constitués à partir de certaines caractéristiques dont:

l'âge:

Tous les répondants âgés de 55 ans et plus (communément étiquetés par la presse populaire comme faisant partie du "pouvoir gris", ont été inclus dans l'échantillon final. Les aînés, pas plus que les autres segments de la population, ne peuvent être considérés comme un ensemble homogène. Il est tout à fait loisible de croire que des tendances particulières se distinguent selon l'avancement de l'âge. Trois sous-groupes ont donc été formés, soit les 55 à 64 ans ou les "jeunes vieux", les 65 à 74 ans et enfin les 75 ans et plus ou les "aînés âgés". Même si on peut qualifier d'arbitraires les limites d'âge choisies, on s'est conformé, dans la présente étude, à celles généralement utilisées

dans des études sur cette population. Ceci a d'ailleurs faciliter la comparaison des résultats avec ceux d'autres études.

le revenu:

La question sur le revenu exigeait que le répondant situe le revenu de son ménage dans une des cinq catégories proposées. Lors des analyses subséquentes, il y a eu refonte en trois catégories, soit moins de 10 000 $, de 10 000 $ à 19 999 $ et 20 000 $ ou plus.

le lieu de résidence:

Afin d'évaluer si le degré d'urbanisation pouvait afffecter les comportements alimentaires, les répondants ont été répartis en deux groupes: d'une part, ceux qui résident en milieu urbain et d'autre part, ceux qui habitent dans des villages, des paroisses et des lieux dits sans désignation officielle.

et le sexe.

A partir des informations recueillies, une brochure destinée aux aînés a été élaborée par le Ministère. Les principaux thèmes de la brochure sont l'alimentation normale et modifiée, des conseils pour la préparation et les achats des aliments ainsi que des menus et recettes adaptés aux besoins des gens âgés.

Conclusion

Ces trois études partagent une approche "douce" à la méthodologie des enquêtes, conditionnée par la capacité de collaboration de nos partenaires. Les études étant fondamentalement descriptives, les données sont de nature qualitative, avec les statistiques appropriées. Cette approche convient au contexte de travail et a donné des résultats fort satisfaisants, tant au niveau de la participation que des retombées. Tant dans la cueillette et la diffusion des données que dans le développement des programmes découlant de ses enquêtes, le MAPA a toujours privilégié le partenariat avec des organismes externes. Tous ces collaborateurs ainsi que de nombreux intervenants des milieux de la santé, de l'éducation et des secteurs privés du monde bio-alimentaire ont su tirer profit des informations issues de ces enquêtes. Encouragé par le succès de ses interventions dans ce domaine, le MAPA entend poursuivre ses activités pour développer de bonnes habitudes alimentaires chez l'ensemble de la population, tout en encourageant la consommation des excellents produits bio-alimentaires québécois.

RÉFÉRENCES

Canada, Santé et Bien-être social. (1977). ***Nutrition Canada. Rapport sur les habitudes alimentaires.*** *Ottawa: Bureau des sciences de la nutrition.*

Québec, ministère des Affaires municipales. (1982). ***Répertoire des municipalités du Québec 1981.*** *Québec: Direction des communications.*

Québec, ministère de l'Éducation. (1983). ***Répertoire des organismes et des écoles 1981-1982.*** *Québec: Direction des études économiques et démographiques.*

Québec, ministère de l'Agriculture, des Pêcheries et de l'Alimentation. (1985). ***Les habitudes alimentaires des ménages québécois. Rapport d'enquête.*** *Québec: Direction des politiques alimentaires.*

Québec, ministère de l'Agriculture, des Pêcheries et de l'Alimentation. (1985). ***Les habitudes alimentaires des jeunes Québécois. Rapport d'enquête.*** *Québec: Direction des politiques alimentaires.*

Québec, ministère de l'Agriculture, des Pêcheries et de l'Alimentation. (1989). ***Les habitudes alimentaires des aînés québécois. Rapport d'enquête.*** *Québec: Service des politiques alimentaires et des relations commerciales.*

Seoane, N.A. (1988). ***Le comportement alimentaire des jeunes Québécois.*** *Rev. can. écon. familiale 38(1):25-29.*

COMPOSANTES SEQUENTIELLES DES RAPPORTS DE DENSITE D'INCIDENCE

MICHEL ALARY, MD, MSc

JACQUES BRISSON, MD, ScD

EQUIPE DE RECHERCHE EN EPIDEMIOLOGIE

DEPARTEMENT DE MEDECINE SOCIALE ET PREVENTIVE

UNIVERSITE LAVAL

QUEBEC

LE 15 MAI 1990

COMPOSANTES SEQUENTIELLES DES RAPPORTS DE DENSITE D'INCIDENCE

INTRODUCTION

1. TAUX D'INCIDENCE = FORCE AVEC LAQUELLE LA POPULATION "NORMALE" S'ECOULE VERS LA POPULATION MALADE.

2. AINSI, SI LA TAILLE DE LA POPULATION "NORMALE" EST VUE COMME UNE FONCTION DU TEMPS, LE TAUX D'INCIDENCE INSTANTANE AU TEMPS T PEUT SE DEFINIR:

   $TI_t = [- d (N_t)/dt] / N_t$, N_t ETANT LA TAILLE DE CETTE POPULATION

3. DENSITE D'INCIDENCE: TAUX MOYEN SUR LA PERIODE T_0-T

   $ID = I/PT$

   I = NOMBRE DE CAS DANS LA PERIODE T_0-T

   PT= PERSONNES-TEMPS A RISQUE CUMULES PAR LA POPULATION OBSERVEE DURANT LA PERIODE

4. CE CONCEPT EST DIFFERENT DU RISQUE QUI SE DEFINIT:

   $R_t = 1 - \exp [ID \times t]$

5. LORSQUE L'ON ETUDIE LA RELATION ENTRE UNE MALADIE ET UN FACTEUR DE RISQUE LE RAPPORT DES DENSITES D'INCIDENCE EST UNE DES MESURES D'ASSOCIATION LES PLUS IMPORTANTES.

6. LES DENSITES D'INCIDENCE PEUVENT SE DECOMPOSER EN PLUSIEURS ETAPES. PAR EXEMPLE, LA DENSITE DE DECES PAR UNE CAUSE QUELCONQUE PEUT SE DECOMPOSER EN DEUX ETAPES: LA DENSITE D'INCIDENCE DE LA MALADIE ET LA DENSITE DE FATALITE CHEZ LES PERSONNES MALADES.

MODELE ET SES HYPOTHESES

DONNEES GLOBALES

	EXP +	EXP -
D+	A	B
PT	N_1	N_0

$IDR = A/N_1 \,/\, \overline{B/N_0}$ = RAPPORT DES DENSITES DE DECES

$ID_1 = A/N_1$ = DECES CHEZ LES EXPOSES

$ID_0 = B/N_0$ = DECES CHEZ LES NON-EXPOSES

DONNEES DE LA PREMIERE ETAPE

	EXP +	EXP -
M+	C	D
PT_1	M_1	M_0

$IDR_1 = C/M_1 \,/\, \overline{D/M_0}$ = RAPPORT DES DENSITES D'INCIDENCE

$ID_{11} = C/M_1$ = INCIDENCE CHEZ LES EXPOSES

$ID_{01} = D/M_0$ = INCIDENCE CHEZ LES NON-EXPOSES

DONNEES DE LA SECONDE ETAPE

	EXP +	EXP -
D+	A	B
PT_2	T_1	T_0

$IDR_2 = A/T_1 \,/\, \overline{B/T_0}$ = RAPPORT DES DENSITES DE FATALITE

$ID_{12} = A/T_1$ = FATALITE CHEZ LES EXPOSES

$ID_{02} = B/T_0$ = FATALITE CHEZ LES NON EXPOSES

POPULATION EN ETAT D'EQUILIBRE:

AUTANT D'INDIVIDUS QUI NAISSENT QUE D'INDIVIDUS QUI DEVIENNENT MALADES QUE D'INDIVIDUS QUI MEURENT OU GUERISSENT.

$$A = C$$

$$B = D$$

$$N_1 = M_1 + T_1$$

$$N_0 = M_0 + T_0$$

DANS CES CONDITIONS, MORRISON (1979) A DEMONTRE LES RELATIONS SUIVANTES:

$$ID = ID_2 \times [ID_1/(ID_1 + ID_2)]$$

DONC,

$$IDR = \frac{ID_{12} \times [ID_{11}/(ID_{11} + ID_{12})]}{ID_{02} \times [ID_{01}/(ID_{01} + ID_{02})]}$$

DANS LE CONTEXTE DES ETUDES CAS TEMOINS BASEES SUR DES CAS INCIDENTS, IL EST POSSIBLE D'ESTIMER LE IDR MAIS LES ID NE PEUVENT ETRE ESTIMES EN TANT QUE TEL:

	EXP	
	+	-
CAS	a	b
TEMOINS	c	d

L'ENSEMBLE DES TEMOINS (b + d) PEUT ETRE VU COMME UN ECHANTILLON DES PERSONNES TEMPS. DANS CE CAS,

$$IDR = OR = ad/bc$$

MAIS ID_1 ne a/c ET ID_0 ne b/d.

POUR ETABLIR LE MEME TYPE DE RELATIONS DANS LES COMPOSANTES SEQUENTIELLES, IL FAUT DONC ELIMINER LES ID DES FORMULES.

EN PARTANT DE LA RELATION QUI INCLUT LES ID:

$$IDR = \frac{ID_{12} \times [ID_{11}/(ID_{11} + ID_{12})]}{ID_{02} \times [ID_{01}/(ID_{01} + ID_{02})]}$$

ON OBTIENT:

$$IDR = IDR_2 \times IDR_1 \times \frac{(ID_{01} + ID_{02})}{(ID_{11} + ID_{12})}$$

$$IDR = IDR_2 \times IDR_1 \times \frac{[(D/M_0) + (B/T_0)]}{[(C/M_1) + (A/T_1)]}$$

EN SACHANT QUE, SELON LES HYPOTHESES DE BASE:

$$A = C$$

$$B = D$$

$$N_1 = M_1 + T_1$$

$$N_0 = M_0 + T_0$$

ON OBTIENT, APRES MANIPULATIONS:

$$IDR = \frac{IDR_1 \times IDR_2}{IDR_1 (T_0/N_0) + IDR_2 (1 - T_0/N_0)}$$

DANS LE CONTEXTE DES ETUDES CAS TEMOINS, T_0/N_0 REPRESENTE LA PROPORTION DE TEMOINS NON EXPOSES QUI ONT FRANCHI LA PREMIERE ETAPE DU PROCESSUS.

PRENONS L'EXEMPLE SUIVANT:

$ID_{01}=0.2$ $ID_{02}=0.2$ $ID_{11}=2.0$ $ID_{12}=1.0$

SELON MORRISON,

$ID_0 = ID_{02} \times [ID_{01} / (ID_{01} + ID_{02})] = 0.1$

$ID_1 = ID_{12} \times [ID_{11} / (ID_{11} + ID_{12})] = 0.67$

$IDR = ID_1 / ID_0 = 6.7$

AVEC NOTRE FORMULE

$$IDR = \frac{IDR_1 \times IDR_2}{IDR_1 (T_0/N_0) + IDR_2 (1 - T_0/N_0)}$$

AVEC $IDR_1 = 10$ $IDR_2 = 5$ $T_0/N_0 = 0.5$

$$IDR = (10 \times 5)/[(10 \times 0.5)+(5 \times 0.5)] = 6.7$$

ILLUSTRATIONS NUMERIQUES

1. EN GENERAL, T_0/N_0 EST LE PARAMETRE LE PLUS INCERTAIN DANS LA RELATION QUE NOUS AVONS ETABLIE. SI LES TEMOINS CONSTITUENT UN VERITABLE ECHANTILLON ALEATOIRE DES PERSONNES-TEMPS A RISQUE, IL N'Y AURA AUCUN PROBLEME. SINON CERTAINES INEXACITUDES PEUVENT SURVENIR.

2. LE PLUS SOUVENT, LA CUEILLETTE DE DONNEES AURA PERMIS D'ESTIMER IDR ET UNE DE SES DEUX COMPOSANTES.

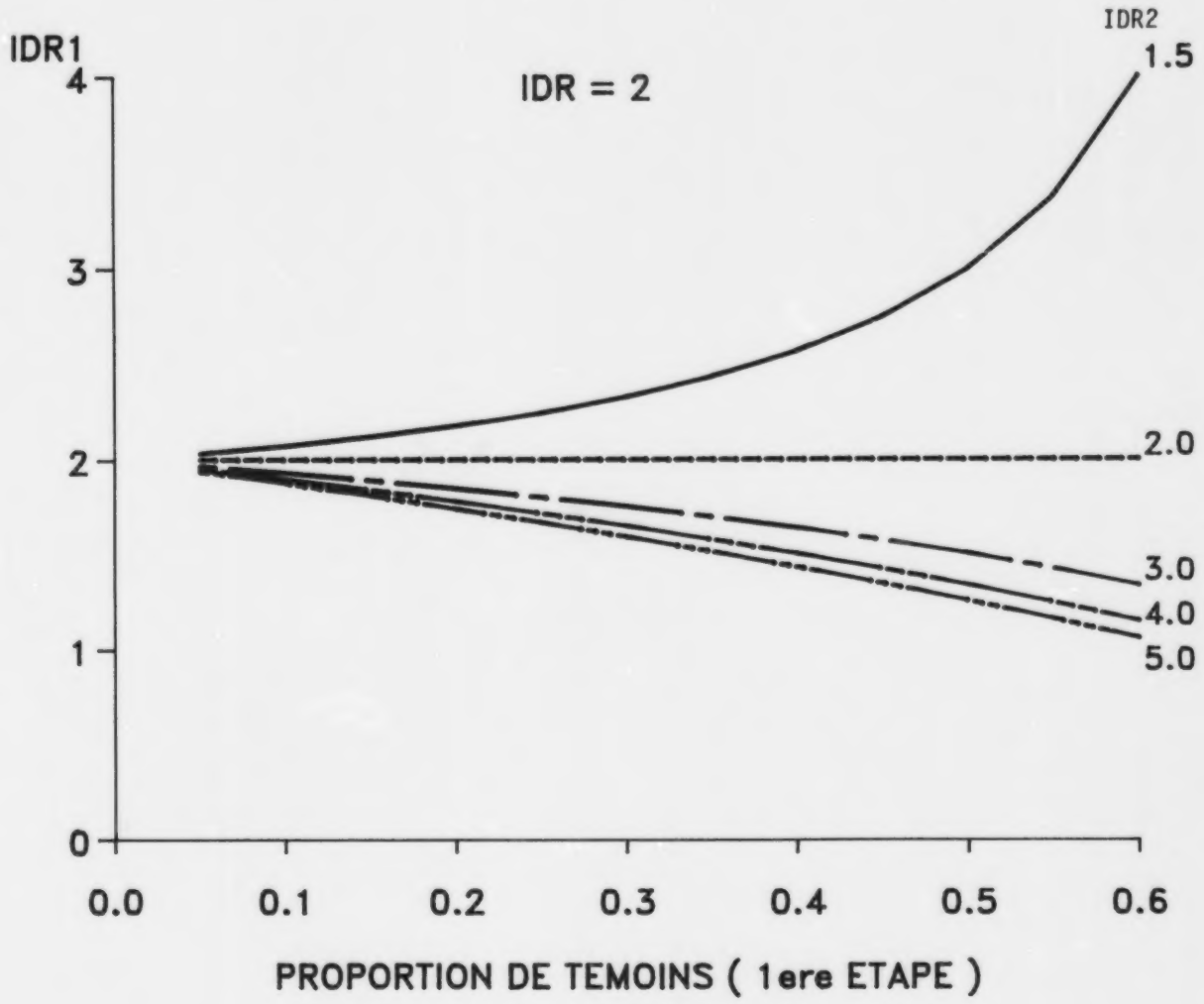
3. LES GRAPHIQUES DES PAGES SUIVANTES PRESENTENT DONC LA VARIATION DE IDR_1 POUR DES IDR ET IDR_2 FIXES EN FONCTION DE CHANGEMENTS DANS T_0/N_0. LA RELATION UTILISEE EST LA SUIVANTE:

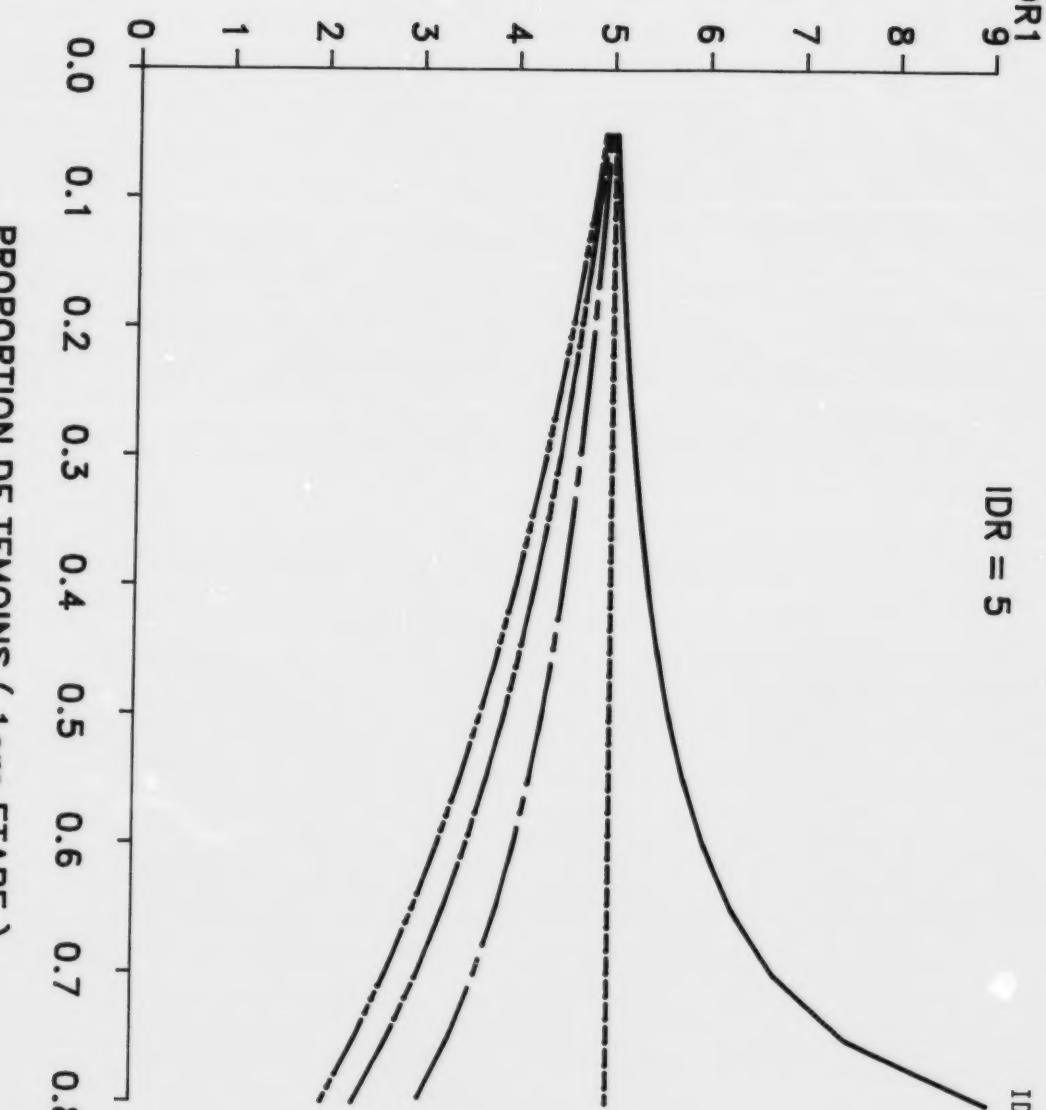
$$IDR_1 = \frac{IDR \times IDR_2 (1 - T_0/N_0)}{IDR_2 - IDR (T_0/N_0)}$$

CONCLUSIONS

1. NOUS AVONS, DANS LE CONTEXTE D'UNE POPULATION EN ETAT D'EQUILIBRE, REUSSI A ETABLIR UNE RELATION ENTRE LES COMPOSANTES SEQUENTIELLES DU IDR D'UNE FACON INDEPENDANTE DES ID EUX-MEMES.

2. CE MODELE POURRA TROUVER CERTAINES APPLICATIONS PRATIQUES DANS LES ETUDES CAS TEMOINS QUI ESSAIENT D'ISOLER L'EFFET D'UN FACTEUR DE RISQUE SUR LES DIFFERENTES ETAPES DU PROCESSUS PATHOGENIQUE.


IDR1
IDR2
IDR = 2
1.5
2.0
3.0
4.0
5.0
4
3
2
1
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
PROPORTION DE TEMOINS (1ere ETAPE)


IDR1
9
8
7
6
5
4
3
2
1
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
PROPORTION DE TEMOINS (1ere ETAPE)
IDR = 5
IDR2
4.5
5.0
6.0
7.0
8.0

PRESENTATION A L'ACFAS
MAI 1990

Titre Calcul des tailles d'échantillons pour la détection d'un rapport de risques, suivant un niveau de confiance et une puissance déterminée

Auteurs Bernard, P-M
Lapointe, C
Département de médecine sociale et préventive,
Faculté de Médecine, Université Laval,
Québec, G1K 7P4,

Introduction

La présentation s'intéresse à l'élaboration de formules pour le calcul de tailles d'échantillons nécessaires à la détection d'un rapport de risque, compte tenu du niveau de signification (α) choisi et de la puissance ($1-\beta$) retenue pour le test statistique. On décrit d'abord deux principes de base sur lesquels doit reposer l'élaboration de telles formules. Puis, on décrit une formule générale respectant ces deux principes. Ensuite, on discute plus spécifiquement de l'utilisation de la formule de Casagrande [1] pour le calcul des tailles d'échantillons dans la détection d'un rapport de cotes [2] tout en présentant la formule issue de la formule générale élaborée. On termine avec une comparaison numérique entre la formule de Casagrande et celle proposée.

DEUX PRINCIPES DE BASE

PRINCIPE I: **Le mode de calcul d'une taille d'échantillon ou d'une puissance devrait concorder avec le mode d'analyse**

Dans la pratique, ce principe est souvent violé. Par exemple, la mesure impliquée dans le calcul de la taille de l'échantillon n'est pas la même que celle utilisée au niveau de l'analyse. Ainsi, pour l'estimation d'un rapport de cotes on base la formule sur une différence de proportion (p_1-p_0 versus RC)[(1)]. On peut souligner d'autres incohérences comme le fait de définir des formules de calcul exact alors qu'au niveau de l'analyse les

tests sont approximatifs; le fait de baser le calcul de taille d'échantillon sur une distribution non conditionnelle alors qu'au niveau de l'analyse le test statistique est conditionnel.

PRINCIPE II: **Le calcul de taille d'échantillon est une procédure a priori**

Les données pour les calculs de tailles d'échantillons sont générées par des hypothèses de recherche ou relèvent de décisions arbitraires du chercheur. Celui-ci, en accord avec ses hypothèses de recherche, peut s'intéresser à la détection d'une certaine différence p_1-p_0 ou d'un rapport de cotes RC suivant un degré de signification α et une puissance $1-\beta$ déterminés. Enfin et ce qui nous occupe principalement ici, le calcul de la taille d'échantillon requiert la connaissance (a priori) de la variance de la mesure en cause. On doit donc admettre que le calcul de la variance de la mesure se fait a priori, dans les limites des hypothèses statistiques en cause et non a posteriori, c'est-à-dire à partir de données que l'on ne connaît pas. Par exemple, pour la différence postulée $D = P_1-P_0$ entre deux proportions, la variance de la statistique d sous l'hypothèse H_0 que D est nulle est donnée par

$$\underline{\text{a priori}} \qquad\qquad \underline{\text{a posteriori}}$$

$$\frac{(k+1)\ P_0(1-P_0)}{k \qquad n_1} \qquad\qquad \frac{p_1q_1 + k(p_1q_0+p_0q_1) + k^2p_0q_0}{n_1k(1+k)}$$

où P_0 désigne un paramètre alors que p_i (i= 0 ou 1) désigne une

valeur observée. On remarque alors que l'expression de la variance varie en fonction du point de vue choisi.

FORMULE DE BASE POUR LE CALCUL APPROXIMATIF D'UNE TAILLE D'ECHANTILLON

En accord avec les deux principes énoncés précédemment, on peut déduire la formule de base suivante pour le calcul d'une taille d'échantillon nécessaire à la détection d'une différence X suivant un niveau de signification α et une puissance $1-\beta$:

$$E_1(X) - E_0(X) \geq Z_\alpha\sqrt{var_0(X)} + Z_\beta\sqrt{var_1(X)} \qquad [1]$$

où $E_1(X)$ (resp. $E_0(X)$) désigne la moyenne de X sous l'hypothèse H_1 (resp. H_0).
Concrètement, X peut désigner une différence de proportions ou de taux, ou encore le logarithme d'un rapport de risques (RR) ou de cotes (RC).Souvent, $E_0(X) = 0$. Par exemple, sous H_0 la différence $P_1-P_0 = 0$; de même, RR (ou RC) est égal à 1 et donc lnRR (ou lnRC) = 0.

De la formule [1], on explicite les valeurs n_1 et n_0 qui servent à déterminer les variances, ce qui est plus facile si on a déjà une relation entre n_1 et n_0: $n_0/n_1 = k$.

PROPOSITION D'UNE FORMULE POUR LE RC

La formule générale [1] appliquée au lnRC (RC noté ici simplement R) se réduit à l'expression suivante:

$$n_1 = [Z_\alpha \sqrt{A_0} + Z_\beta \sqrt{A_1}]^2/(\ln R)^2$$

où $A_0 = n_1 var_0 \ln R = (k+1)/kp_0q_0$

et $A_1 = n_1 var_1 \ln R = (k+1)/kp_0q_0 + (R-1)[p_0^2(R-1)+2p_0-1]$

Sur cette formule proposée, on peut faire les remarques suivantes:

1. Cette formule est basée sur la transformation logarithmique du RC (lnRC)
2. La variance utilisée est celle de lnRC; elle est estimée dans l'approche a priori par la méthode des séries de Taylor.
3. Le calcul est cohérent avec l'analyse qui utilise la transformation ln du RC, par exemple avec la régression logistique univariée.

FORMULE DE CASAGRANDE POUR LE RC

La formule de Casagrande et al[2] est celle-ci:

$$n_1 = A.W/k(p_1-p_0)^2$$

où $A = [Z_\alpha \sqrt{(k+1)pq} + Z_\beta \sqrt{kp_1q_1 + p_0q_0}]^2$

$W = \frac{1}{4}[(1 + \sqrt{1 + 2(k+1)(p_1-p_0)/A}]^2$

$p = 1 - q = (p_1+kp_0)/(1+k)$

$R = p_1q_0/p_0q_1$ (rapport de cotes)

Sur cette formule de Casagrande, on peut faire les remarques suivantes:

1. Cette formule est celle de Schlesselman, avec correction de continuité élaborée par Casagrande, généralisée par Fleiss à la situation des échantillons inégaux.
2. Elle est souvent utilisée pour le calcul de tailles d'échantillons nécessaires à la détection d'un rapport de cotes bien qu'elle soit construite en considérant une différence de proportions
3. Les variances en cause sont celles de différences de proportions, variances estimées a posteriori.

Ainsi, pour l'estimation d'un rapport de cotes, la formule de Casagrande se trouve à violer les deux principes énoncés précédemment.

COMPARAISON NUMERIQUE

Nous complétons la présentation avec quelques comparaisons numériques entre des tailles d'échantillons calculées sous les mêmes conditions par chacune des deux méthodes discutées ici. Dans chacun des tableaux, le nombre de la première ligne est la taille calculée par la formule proposée et celui de la seconde par la formule de Casagrande.

A) Comparaison numérique pour la détection d'un rapport de cotes RC, suivant différentes valeurs de p_0 et k, pour $\alpha = 0,05$ et $1-\beta = 0,80$.

	p_0									
	0,01	0,05	0,10	0,20	0,25	0,30	0,40	0,50	0,60	0,80
RC=1,5										
k=1	6672	1415	763	448	390	356	325	325	352	571
	7157	1500	797	454	390	351	311	303	320	494
k=2	4901	1041	563	332	290	265	243	244	265	434
	5256	1104	588	337	290	262	233	228	242	376
k=4	4009	854	462	274	240	219	202	203	222	365
	4303	905	483	278	240	217	194	190	203	317
RC=2,0										
k=1	2087	449	246	150	133	123	115	119	133	226
	2372	499	266	153	132	120	107	105	113	177
k=2	1512	327	180	110	98	91	86	89	100	173
	1719	363	195	113	98	89	80	80	86	137
k=4	1220	264	146	90	81	75	72	75	84	146
	1391	295	159	93	81	74	67	67	72	117
RC=3,0										
k=1	732	161	91	58	53	50	49	53	61	112
	913	193	104	61	53	48	44	44	47	77
k=2	521	116	66	43	39	37	37	40	47	86
	651	139	75	45	39	36	33	33	37	61
k=4	412	92	53	35	32	31	31	33	39	73
	519	111	61	37	32	30	28	28	31	52

B) Comparaison numérique pour la détection d'un rapport de cotes R, suivant différentes valeurs de p_0 et k, pour $\alpha = 0,05$ et $1-\beta = 0,95$.

	p_0									
RC=1,5	0,01	0,05	0,10	0,20	0,25	0,30	0,40	0,50	0,60	0,80
k=1	11381	2413	1301	763	665	606	553	553	600	974
	12191	2561	1365	783	675	609	544	533	566	885
k=2	8527	1808	975	572	498	454	414	414	449	730
	8857	1866	999	577	499	452	406	401	429	679
k=4	7089	1503	811	476	415	378	345	345	374	607
	7186	1518	815	474	412	374	338	335	361	576

A partir de ces exemples numériques, on remarque que, relativement aux tailles d'échantillons obtenues par la formule de Casagrande, celles calculées par la formule proposée sont d'autant plus grandes que la fréquence p_0 de la caractéristique est plus petite que 0,25. Et vice versa pour p_0 plus grand que 0,25.

Bibliographie

(1) Breslow N.E., Day N.E.: Statistical Methods in Cancer Research, Volume II - The design and Analysis of Cohort Studies. IARC Scientific Publications No. 92, 1987, 289.

(2) Casagrande J.T., Pike M.C., Smith P.G.: An improved approximate formula for calculating sample sizes for comparing two binomial distributions. Biometrics, 1978, 34, 483-486.

UTILISATION DES MODELES DE REGRESSION DE COX DANS L'EVALUATION DU TAUX DE SURVIE D'UNE GREFFE DE CORNEE.

Julie Bernier, M.Sc., Hélène Marie Boisjoly, M.D., Paul-Marie Bernard, M. Sc..

L'analyse des durées de vie ou données de survie est une technique utilisée fréquemment en biologie et en médecine. En ophtalmologie, cette technique est particulièrement utile pour la modélisation du phénomène de rejet des greffes de cornée. Le département d'ophtalmologie du CHUL possède une banque de données contenant des renseignements sur plus de 500 patients ayant subi une greffe de cornée et complété par la suite un suivi d'une durée moyenne de 2.5 ans. Voyons l'application du modèle de régression de Cox et sa validation sur un exemple tiré de cette banque de données.

1- Données de survie.

Un modèle de données de survie en est un où on s'intéresse à une variable aléatoire T qui est un temps de vie, c'est-à-dire le moment auquel se produit un événement particulier. On peut modéliser

- f(t), la fonction de densité de probabilité de T
- F(t), sa fonction de répartition
- S(t), la fonction de survie définie comme $P(T \geq t)$
- $\lambda(t)$, le taux de mortalité ou de défaillance défini comme $f(t)/S(t)$.

Il existe diverses méthodes pour estimer ces fonctions ou simplement certains de leurs paramètres. On travaille habituellement aussi avec un ensemble de covariables et on note alors les fonctions $f(t \mid \mathbf{x})$, $F(t \mid \mathbf{x})$, $S(t \mid \mathbf{x})$, $\lambda(t \mid \mathbf{x})$.

Généralement, les données utilisées sont des données censurées, ce qui signifie que pour chaque individu on a, outre les covariables, la paire de variables suivante: (Ti,ci). La première variable représente soit le temps de vie soit la censure, la deuxième variable est la variable indicatrice de la censure. Dans une expérience médicale dont la durée est fixée, si l'entrée d'un individu dans l'expérience se fait de façon aléatoire on parle de censure aléatoire.

2- Régression de Cox.

Lorsqu'on fait l'étude de données de survie avec un ensemble de covariables, on s'intéresse à la distribution conditionnelle de T étant donné un vecteur **x**. On peut alors utiliser un modèle de régression. Cox (1972) propose une approche non paramétrique à ce problème: **Le modèle à taux de mortalité proportionnel**. Ce modèle est construit sous l'hypothèse

que les covariables ont des effets multiplicatifs sur le taux de mortalité $\lambda(t \mid \mathbf{x})$. Cette hypothèse s'écrit:

$$\lambda(t \mid \mathbf{x}) = \lambda_0(t) \exp(\mathbf{x}'\beta)$$ où la fonction $\lambda_0(t)$ est arbitaire et non spécifiée.

Sous cette hypothèse, le rapport des taux de mortalité chez des individus ayant des ensembles de covariables différents $\mathbf{x_1}$ et $\mathbf{x_2}$ est:

$$\lambda(t \mid \mathbf{x_1})/\lambda(t \mid \mathbf{x_2}) = \exp((\mathbf{x_1}-\mathbf{x_2})'\beta)$$ qui ne dépend pas de t.

On appelle ce rapport le risque relatif des deux individus et on le note $RR(\mathbf{x_1},\mathbf{x_2})$.

Cox a proposé d'estimer β à l'aide de sa fonction de vraisemblance et ce sans faire d'hypothèse sur $\lambda_0(t)$. C'est ce qu'on appelle **la régression de Cox**. Le vecteur β qui maximise la fonction de vraisemblance s'obtient à l'aide du processus itératif de Newton-Raphson. L'estimateur ainsi construit est asymptotiquement Normal et il est possible de faire un test d'hypothèse sur les β_i du modèle. Ceci permet de déterminer quelles sont les variables qui ont un effet significatif sur le taux de mortalité.

3- Estimation de la fonction de survie.

Le modèle de Cox possède en fait deux inconnus; le vecteur des coefficients de régressions β et le taux de mortalité de base $\lambda_0(t)$. En général, les cliniciens sont intéressés non seulement à savoir quelles sont les variables influentes mais aussi à prédire la probabilité de survie de la greffe après une période de temps t. Pour faire ce genre de prédiction on peut procéder de deux façons. On peut premièrement ignorer les covariables et calculer une courbe de survie estimée $S(t)$. Cette estimation se fait à l'aide de la méthode de Kaplan-Meier (1958), la méthode consiste à estimer la survie en un point à partir d'un produit d'estimations de proportions faites aux temps précédents. Evidemment on apporte une correction pour tenir compte des censures.

La deuxième facon de procéder tient compte des covariables. On doit, après avoir estimé β, estimer le taux de mortalité de base $\lambda_0(t)$ ou de manière équivalente la fonction de survie de base $S_0(t)$. En effet ces deux fonctions sont reliées par l'équation

$$S_0(t) = \exp\left(-\int_0^t \lambda_0(u)\, du\right).$$

Et la fonction de survie de T sachant $\mathbf{x}$ peut s'écrire comme

$$S(t \mid \mathbf{x}) = \exp\left(-\int_0^t \lambda(u \mid \mathbf{x})\, du\right).$$

Ou plus simplement on peut se servir de β et de $S_0(t)$ pour calculer la fonction de survie

$$S(t \mid \mathbf{x}) = [S_0(t)]^{\exp(\mathbf{x}'\beta)}.$$

Pour estimer la fonction de survie de base $S_0(t)$, on utilise une approche non paramétrique. Après avoir calculé β on estime $S_0(t)$ à l'aide d'une technique analogue à celle de Kaplan-Meier. Cette méthode est explicité dans Lawless (1982). $S_0(t)$ et $S(t \mid \mathbf{x})$ obtenues par cette méthode sont bien entendu des fonctions décroissantes en escalier.

4- Validation du modèle.

Dans le cas du modèle de Cox, la validation consiste à vérifier que l'hypothèse de proportonnalité est applicable à un ensemble de données particulier. Deux méthodes graphiques sont proposées par Cox et Snell (1968) pour vérifier cette hypothèse. La première consiste à faire l'étude des résidus définis comme:

$$r_i = \lambda(t_i \mid x_i) = -\ln(S(t_i \mid x_i)).$$

Avec cette définition, l'ensemble des résidus en est un qui contient des valeurs censurées. Si l'hypothèse de proportonnalité est vérifiée, cet ensemble peut être considéré comme issu d'un échantillon aléatoire d'une distribution exponentielle de moyenne 1. A ce moment, si on calcule $S_1(e)$ l'estimateur de Kaplan-Meier de la fonction de survie des e_i, le graphique de $\ln(S_1(e))$ en fonction de e devrait nous montrer une relation linéaire de pente -1.

Cependant, il faut utiliser cette méthode avec prudence car, dans le cas où aucune forme précise n'est supposée pour $S_0(t)$, la distribution des résidus peut différer substantiellement de la loi exponentielle (Lagakos,1981).

La deuxième méthode utilise la même définition des résidus. Lorsque les résidus sont censurés, on peut montrer que $E(r_i \mid r_i \geq a_i^*)=a_i^*+1$. Pour les observations t_i censurées, un ajustement raisonnable serait donc de prendre

$$r_i^* = (-\ln(S_0(t_i)) \exp(x'\beta) + 1.$$

Ceci permet de traiter l'ensemble des résidus comme un ensemble non-censuré et une droite de Henry sur ces résidus nous donne une validation graphique du modèle.

Il existe aussi des méthodes de validation avec graphique et test qui permettent d'étudier une à une les covariables et leur effet sur le modèle. Une d'elles consiste à estimer et comparer les courbes de survie de base dans diverses strates des covariables. Cette méthode demande beaucoup de données réparties équitablement dans chacune des strates. Une autre méthode proposée par Blossfeld, Hamerle et Mayer (1989) consiste à introduire dans la régression des variables qui dépendent à la fois du temps et des covariables et dont les β_i sont nuls si l'hypothèse est respectée. Cette méthode est malheureusement fort couteuse, on peut donc choisir de l'utiliser seulement si les méthodes globales vues précédemment laissent supposer que l'hypothèse de proportionnalité n'est pas respectée.

5- Exemple

Pour illustrer ces méthodes, voici les résultats obtenus pour un échantillon de 438 patients ayant subi une greffe de cornée au CHUL entre 1986 et 1989. Les patients sont tous agés de plus de quinze ans et ceux ayant subi de multiples greffes ne sont inclus dans l'échantillon qu'une seule fois. La variable dépendante est l'apparition d'un épisode de rejet . Cette variable contient une censure aléatoire. Les covariables qui nous intéressent sont la dimension du greffon, la vascularisation du receveur, la compatibilité tissulaire HLA-A, B le diagnostique qui sont des variables dichotomiques (les individus sont classés à risque ou non), et finalement l'âge du receveur qui est une variable continue.

Pour commencer on a calculé la fonction de survie indépendemment des covariables, elle est illustrée à la figure 1. Cette fonction, qui est une estimation par la méthode de Kaplan-Meier, nous permet de constater que pour l'ensemble de la population la survie estimée à 12 mois est de 91% et que la survie estimée à 24 mois est de 80%. Si on veut évaluer la survie pour certains

groupes plus particuliers, on peut faire une régression de Cox multiple avec les cinq covariables suggérées. Voici le résultat obtenu:

variable	β	p-value	RR
dimension	.45	.037	2.5
vascularisation	.59	.018	2.1
HLA-A,B	.71	.001	2.9
diagnostique	.30	.22	----
âge	-.002	.67	----

Pour une variable dichotomique, le risque relatif (RR) représente

$$\frac{\lambda\,(t \mid X1=1 \text{ et les autres variables au niveau moyen})}{\lambda\,(t \mid X1=0 \text{ et les autres variables au niveau moyen})} = \exp(\beta i)$$

Si on veut calculer le risque relatif pour des individus ayant plusieurs variables à des niveaux différents, il est nécessaire d'introduire des termes d'interaction au modèle de régression. Il est aussi intéressant, pour les variables dichotomiques qui ont un effet significatif sur la survie de la greffe, de comparer les courbes de survie pour les deux catégories de la variable. Le graphique des courbes de survie pour chaque niveau de la variable HLA-A, B est représenté à la figure 2. Ces courbes permettent d'estimer, entre autres, que, chez les patients bien appariés, la survie à 24 mois est de 87% tandis qu'elle n'est que de 70% chez ceux mal appariés. La régression quand à elle nous indique que le taux de mortalité (ici, le taux de rejet) est environ 2.9 fois plus élevé chez les patients mal appariés. On peut de plus si nécessaire construire un intervalle de confiance pour le risque relatif en construisant préalablement un intervalle pour β.

Finalement, pour valider de façon globale notre modèle de régression on peut utiliser la droite de Henry ou la fonction de survie estimée des résidus, ces deux graphiques sont représentés respectivement aux figures 3 et 4. Les résultats obtenus par ces deux méthodes sont très satisfaisants, il n'y a donc pas lieu de faire une étude plus approfondie sur les diverses covariables. On peut en toute confiance utiliser le modèle de Cox pour décrire le comportement du phénomène de rejet des greffes de cornée en fonction des covariables choisies pour cet exemple.

Fig.1 : Survie estimée

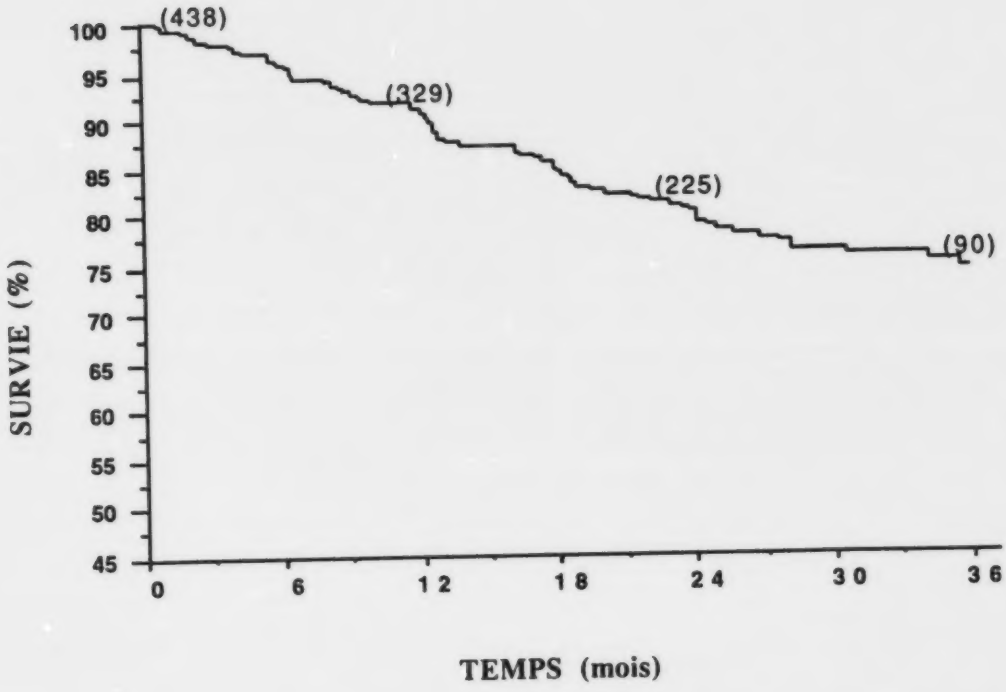


Fig.2 : Survie estimée selon la compatibilité HLA-A,B

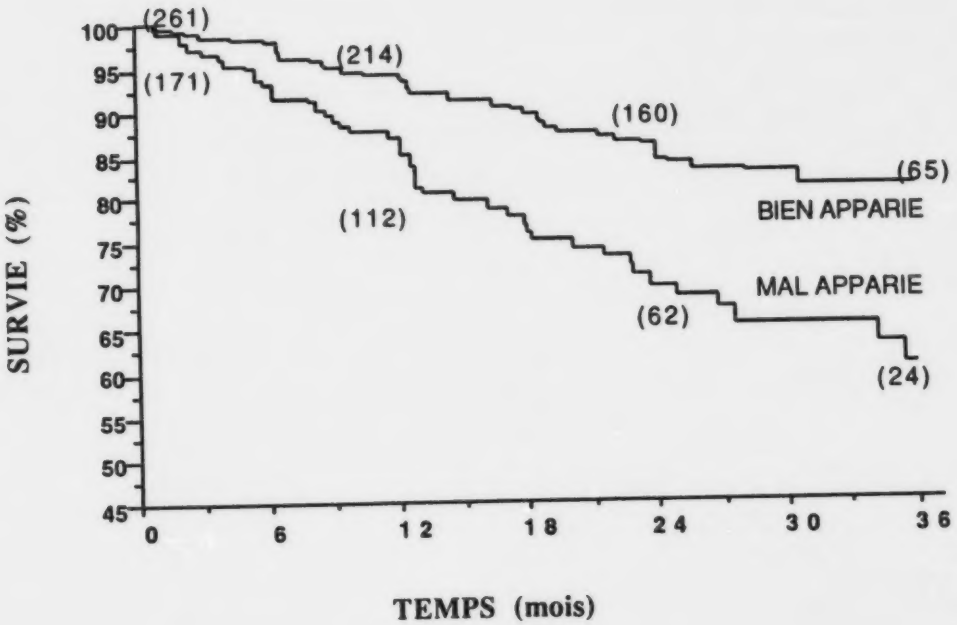


Fig.3 : Droite de Henry

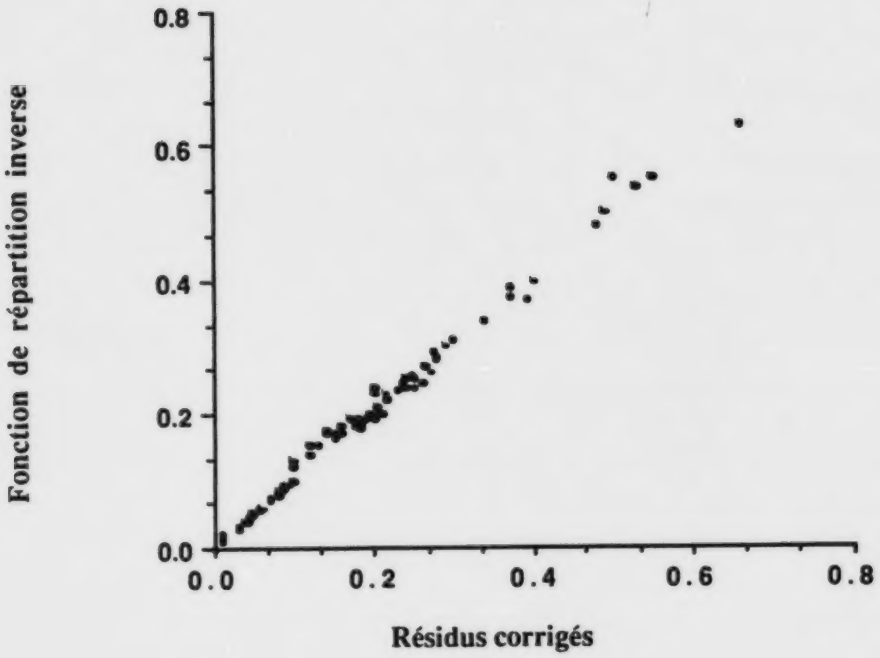
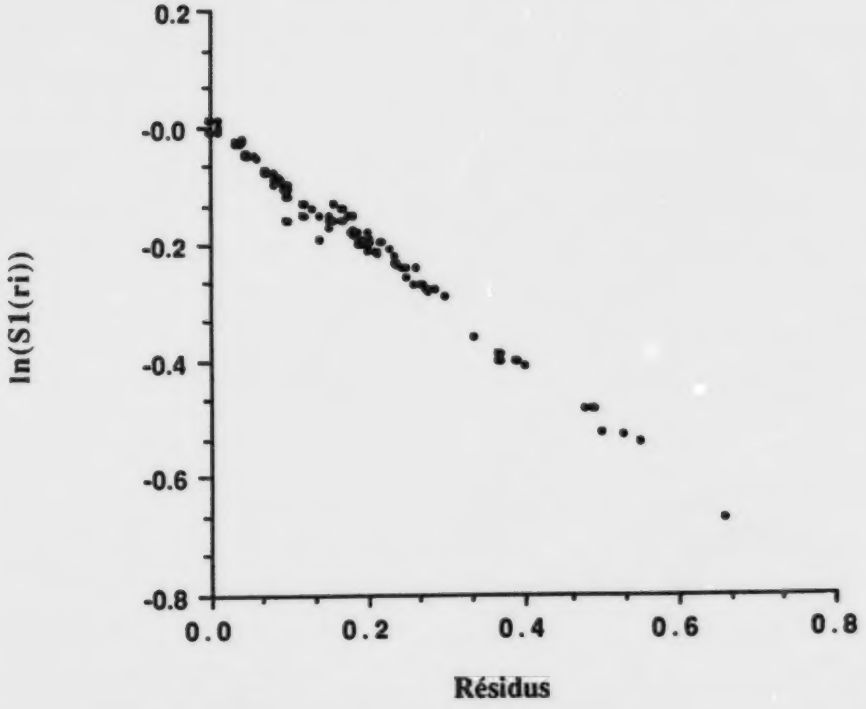


Fig.4 : Méthode de la survie estimée



BIBLIOGRAPHIE

Allison, P.D. (1984). *Event History Analysis. Regression for longitudinal Event Data.* Beverly Hills, California: Sage.

Blossfeld, H-P, Hamerle, A. et Mayer, K.U. (1989). *Event History Analysis.*Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics* , 30, 89-100.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, B, 34, 187-220.

Cox, D.R. (1975).Partial likelihood. *Biometrika*, 62, 269-276.

Cox, D.R. et Snell, E.J. (1968). A general definition of residuals (with discussion). *Journal of the Royal Statistical Society*, B, 30, 248-275.

Kaplan. E.L et Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457-481.

Lagakos, S.W. (1981). The graphical evaluation of explanatory variables in proportonnal hazard regressions models. *Biometrika*, 68, 93-98.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data.* New-York: Wiley.

Peto, R. (1972). Contribution to discussion of paper by D.R. Cox. *Journal of the Royal Statistical Society*, B, 34, 205-207.

Comment expliquer les disparités entre les taux de mortalité provenant du ministère fédéral Santé et Bien-être social Canada et ceux du ministère provincial de la Santé et des Services sociaux ?

par Lise Emond et Jean-Pierre Courteau, DSC de l'Outaouais.

(Texte d'une conférence présentée au colloque "Méthodes et domaines d'application de la statistique 1990" du Bureau de la statistique du Québec, au 58e congrès de l'ACFAS, à Québec, le 15 mai 1990)

Introduction

Pour réaliser son mandat de connaissance et de surveillance de l'état de santé de la population de la région 07, le Département de santé communautaire (DSC) de l'Outaouais analyse et interprète les données de mortalité provenant de deux sources: d'une part, de Santé et Bien-être Canada (SBE), en vertu d'une entente spécifique, et d'autre part du Ministère de la Santé et des Services sociaux (MSSS), dans le cadre du "Système permanent de surveillance de l'état de santé de la population".

Or, une comparaison des taux standardisés (par 100,000) fournis par chacune des deux sources, pour les années regroupées de 1979 à 1983, révèle des disparités assez importantes. Comment s'expliquent ces disparités? C'est ce que nous avons voulu savoir dans la présente étude, afin d'arriver ultimement à choisir la meilleure source de données de mortalité pour la surveillance et la planification régionale dans l'Outaouais.

Méthodes

A titre d'exemple, nous avons choisi de reconstituer le calcul des taux standardisés de mortalité **"toutes causes"** du MSSS, pour les hommes et pour les femmes, en modifiant un à un les facteurs pouvant être responsables de l'écart observé avec les taux de SBE Canada. Nous espérons arriver, par ces modifications successives des taux du MSSS, à des taux identiques à ceux fournis par SBE.

Ainsi, à l'origine, les taux standardisés "toutes causes" fournis par les deux organismes étaient les suivants :

Tableau 1 : Taux standardisés de mortalité "toutes causes", par 100 000, région de l'Outaouais, 1979-1983.

	SBE	MSSS
Hommes	**911.7**	**863.4**
Femmes	**478.8**	**579.9**

On note que les taux masculins de SBE sont plus élevés que ceux du MSSS, et que la différence est inversée chez les femmes. Les taux pour la province de Québec, dans les publications respectives des deux ministères, suivent les mêmes tendances. Cette observation est intriguante, et s'applique aussi à la plupart des causes de décès spécifiques, mais pas toutes (ex. suicide, accidents de la route).

Les facteurs que nous considérerons dans la reconstitution du calcul des taux du MSSS seront les suivants :

1. la méthode de standardisation;
2. le choix de la population de référence;
3. le nombre de décès dans la région de l'Outaouais;
4. la population de la région de l'Outaouais.

Résultats

Le premier facteur que nous avons considéré était la methode de standardisation utilisée. Tant Santé et Bien-être que le MSSS utilisent la méthode directe: les taux regionaux dans chaque groupe d'âge sont pondérés selon l'importance de ce groupe d'âge dans la population de référence. Les taux du MSSS sont cependant standardisés jusqu'à 85 ans, alors que ceux de SBE le sont jusqu'à 75 ans. Pour eliminer l'effet de ce facteur, nous avons repris la standardisation des taux du MSSS jusqu'à 75 ans.

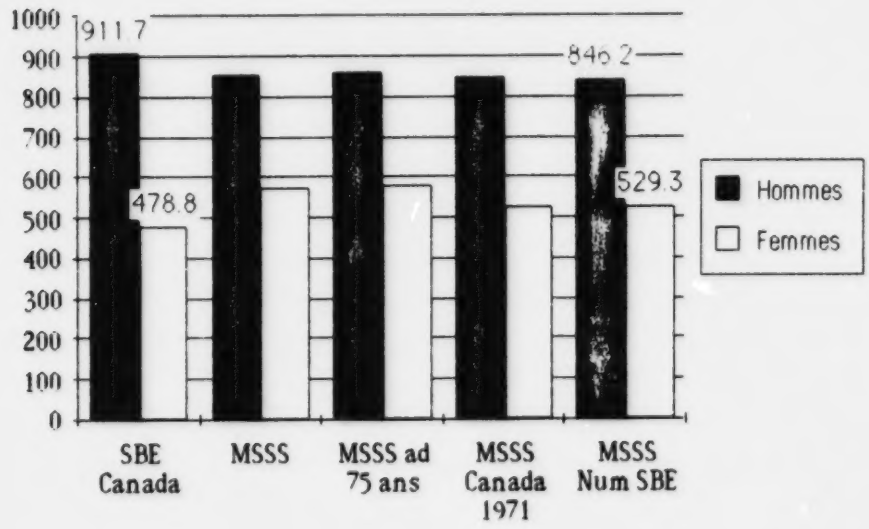
Le second facteur semblait à priori le plus important. La population de reference utilisée par le MSSS était celle du Québec au recensement de 1981, alors que l'organisme fédéral utilisait plutôt la population du Canada en 1971 . Un examen attentif de la structure d'âge des populations du Québec et du Canada montre qu'il s'agit en fait de populations de structure d'âge identique, mais considérées ici à dix ans d'intervalle. Pour estimer l'effet de ce facteur, nous avons repris la standardisation des taux bruts du MSSS, en utilisant la population de référence relativement jeune de Canada 1971.

Le troisième facteur à considérer était le "numérateur" utilisé, c'est-à-dire le nombre de décès dans l'Outaouais pour la période 1979-1983. Pour déterminer ce numérateur, Santé et Bien-être Canada et le MSSS utilisent tous deux un fichier de mortalité informatisé, constitué à partir de certificats de décès compilés à l'échelle provinciale. Le MSSS procède à une allocation des décès par régions administratives selon les codes de municipalités, et à une codification des causes de décès inscrites sur chaque certificat. SBE reprend ce fichier et y effectue quelques modifications mineures. Effectivement, le nombre total de décès attribués à l'Outaouais par chaque organisme était très semblable. Nous avons substitué au numérateur du MSSS celui de SBE Canada aux fins de la présente simulation.

Le dernier facteur que nous avons considéré était la population de la région de l'Outaouais à la mi-période, soit le 30 juin 1981. L'estimation de cette population, et de sa distribution par âge et par sexe, entre dans le calcul des taux bruts de décès par groupes d'âge, qui seront ensuite pondérés dans le processus de standardisation. Nous avons pu vérifier que SBE Canada et le MSSS estimaient tous deux la population de l'Outaouais à partir des données du recensement, et que les deux organismes retenaient les mêmes codes de municipalités, et utilisaient virtuellement les mêmes décomptes et structures d'âge de population pour l'Outaouais. Ce facteur etait donc éliminé d'emblee comme explication potentielle des différences observées entre les taux standardisés.

Les modifications successives des taux du MSSS obtenues par l'exercice de substitution des trois premiers facteurs mentionnés sont illustrées graphiquement à la figure 1.

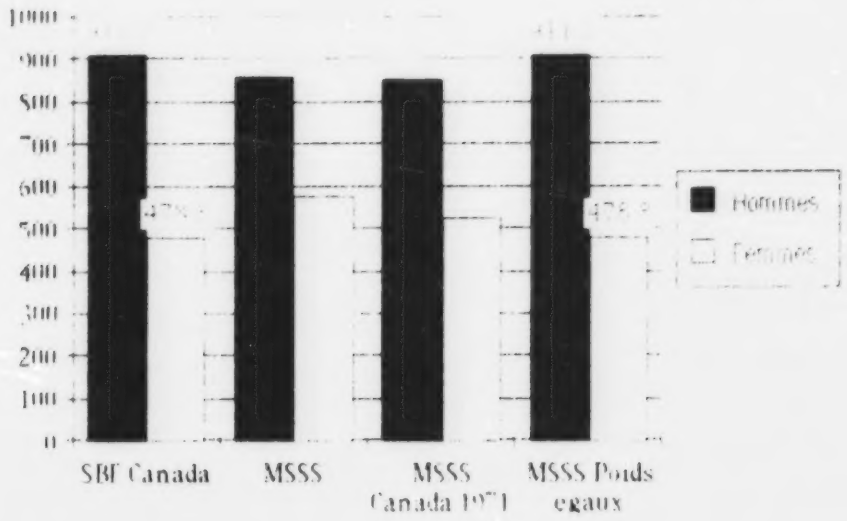
Figure 1 : Modifications successives des taux de mortalité standardisés "toutes causes" du MSSS et comparaison avec les taux de SBE Canada ; région de l'Outaouais ; 1979-1983. Les taux sont exprimés par 100,000.



A notre grande surprise, les taux "modifiés" du MSSS restaient très différents de ceux de SBE. La différence pour les taux masculins s'était même accentuée... Cependant, une discussion avec un statisticien de SBE Canada permettait de mettre en lumière un facteur très important, non explicité dans les notes méthodologiques accompagnant les données: l'organisme fédéral utilise, dans la standardisation des taux masculins et féminins, la structure d'âge de la population de référence **pour les deux sexes réunis.** Cette

particularite, soit l'attribution de **poids égaux** aux taux bruts par groupes d'age des femmes et des hommes, modifie considerablement la mecanique de la standardisation, et produit des taux standardises fort differents. Pour illustrer ceci nous avons repris la standardisation des taux du MSSS, deja modifies par la substitution des facteurs mentionnes precedemment, selon la methode utilisee par SBE Canada. Les nouveaux taux obtenus, appeles "MSSS poids egaux", sont identiques a ceux calcules par SBE Canada (figure 2).

Figure 2 : Effet additionnel sur les taux standardisés modifies du MSSS, de l'utilisation de poids égaux pour les hommes et pour les femmes ; région de l'Outaouais ; 1979-1983. Les taux sont exprimés par 100,000.



Discussion

L'étape suivante consistait à critiquer les deux méthodes et à choisir la plus appropriée aux fins de surveillance et de planification régionale. Un statisticien chevronné de Santé et Bien-être Canada effectuait à notre demande une revue de littérature sur le sujet, et nous fournissait une discussion théorique des différentes mécaniques de standardisation en cause. Nous reprendrons ici quelques points essentiels de cette discussion.

La méthode la plus répandue de standardisation directe des taux est celle qui consiste à utiliser les poids de la population de référence féminine pour les taux féminins, et ceux de la population de référence masculine pour les taux masculins. Instinctivement, cette méthode est plus conforme à un des objectifs de la standardisation, soit de rendre comparables des taux observés au sein d'une population dont la structure d'âge se modifie dans le temps. Par ailleurs, cette méthode ne permet pas de comparer directement les taux standardisés masculins et féminins, puisqu'ils sont calculés à partir de deux populations de référence différentes.

En général, à des fins de planification régionale, le choix d'une population de référence la plus récente possible, et dont la structure d'âge se rapproche le plus de celle de la population à l'étude, est préférable. Ceci dit, la valeur absolue des taux standardisés n'a aucune signification réelle, mais leur valeur relative sert à apprécier des taux provenant de deux populations distinctes dans le temps ou

dans l'espace, pourvu qu'ils soient construits à partir de la même population de référence.

Conclusion

Les facteurs expliquant les disparités ayant été déterminés, et les avantages relatifs de chacun des mécanismes de standardisation étant connus, il restait encore à décider quelle source de données serait utilisée par le DSC de l'Outaouais. Cette décision faisait intervenir des facteurs additionnels d'ordre stratégique et pratique.

En résumé, l'utilisation des données de Santé et Bien-être Canada présente quatre avantages :

1. Les données de SBE couvrent quatre périodes de 5 ans dans l'intervalle de 1969 à 1988. Les données du Ministère sont publiées pour les périodes 1975-1977, 1979-1983 et 1985-1987, donc des périodes de durée variable de trois, cinq et trois années respectivement.

2. Les données de SBE permettent de raffiner l'analyse des taux de mortalité à l'échelle sous-régionale, par territoires de CLSC. SBE utilise les codes de municipalités correspondant aux territoires de CLSC "réels", tels que fournis par le DSC. Le MSSS utilise le découpage territorial déterminé lors de l'implantation des CLSC. Or, ce découpage est maintenant caduc, surtout en régions rurales. Des ententes locales entre CLSC voisins ont permis de respecter davantage les regroupements naturels de population. Par exemple, certaines

portions de territoire ont été restituées à des CLSC auxquels elles étaient reliées préférentiellement à cause de la configuration du réseau routier.

3. Les taux standardisés de SBE sont présentés par groupes d'âge de 1-14 ans, 15-34 ans, 35-69 ans, 69 ans et plus, et pour tous les âges réunis. Ces groupes d'âge sont attrayants en ce qu'ils représentent les différentes étapes de la vie. Les données du MSSS sont présentées par groupes d'âge quinquennaux et pour tous les âges réunis.

4. SBE produit des taux standardisés et des ICM (population de référence: Québec) pour 70 causes spécifiques de décès, et pour tous les groupes d'âge mentionnés. Le MSSS présente les mêmes taux et indices pour 11 causes spécifiques par groupes d'age quinquennaux, et pour 56 causes "tous âges".

Par ailleurs, les données provenant du MSSS sont publiées de façon régulière pour toutes les régions socio-sanitaires, et permettent d'établir des comparaisons entre territoires de DSC. De telles comparaisons sont impossibles avec les données de SBE Canada, puisque celles-ci sont produites sur demande seulement, et à la suite d'ententes spécifiques avec certains Départements de santé communautaire.

Pour toutes ces raisons, nous devons retenir les deux sources de données dans la poursuite de nos objectifs régionaux de connaissance/surveillance. Nous allons cependant privilégier les analyses tirées des données de SBE Canada, notamment dans un portrait de l'évolution de la mortalité dans l'Outaouais de 1969 à 1989

(actuellement en préparation). De plus, nous prévoyons dans les prochains mois doter le DSC d'un centre autonome de traitement des données de mortalité régionales, avec la collaboration et dans la continuité des méthodes de Santé et Bien-être Canada.

REMERCIEMENTS

Les auteurs désirent remercier M. Robert Semenciw, statisticien au LCDC, et M. Russell Wilkins, analyste principal à la Section de la recherche sur l'hygiène du travail et de l'environnement de Santé et Bien-être Canada, qui ont collaboré à la réalisation de ce rapport.

RISQUE DE DIVULGATION
(définition et applications)

par Jean-René Boudreau[1]

1 Introduction

1.1. Une grande variété de règles de confidentialité sont utilisées pour préserver la confidentialité des données rapportées par les unités d'enquêtes. Ces règles sont l'utilisation de classes restreintes, la suppression de cellules, l'arrondissement aléatoire ou déterministe et d'autres méthodes d'addition de bruit aux données. Lorsqu'une règle est utilisée pour un produit de diffusion, sa justification et ses propriétés sont souvent mises en question. La recherche sur la confidentialité doit donc mettre l'accent sur la justification et sur l'optimalité du choix d'un ensemble de règles pour un produit de diffusion donné. Nous donnons dans cet exposé, un nouveau cadre de travail pour évaluer l'efficacité des règles de confidentialité. Avec cette approche, nous serons peut-être en mesure de construire des règles de confidentialité plus objectives qui prendront en considération le degré de risque qu'un Bureau de statistique est prêt à prendre lorsqu'il diffuse de l'information.

2 Notions d'information confidentielle et de risque de divulgation

2.1. Lorsqu'un Bureau de statistique diffuse de l'information sous forme de produits (seuls les produits sous forme de tableaux de fréquences ou de grandeurs sont étudiés), deux sortes d'information sont en réalité dévoilées. Il y a d'abord l'information analytique, c'est-à-dire l'information pertinente aux analyses. L'autre information, que nous appellerons "confidentielle", n'est utile que pour rattacher une valeur donnée à un particulier, une entreprise. L'information analytique est de haut niveau tandis que l'information confidentielle est de bas niveau. L'information confidentielle ne sert en aucune façon aux analyses et il est impossible de l'éliminer complètement.

2.2 Définition. a) L'*information confidentielle* d'un particulier i sur une variable discrète Y pour un produit de diffusion **B** est la fonction de répartition conditionnelle de Y du produit **B** évaluée à i. Le conditionnement est la tribu la plus fine générée par les variables de **B** à l'exception de Y.

---

1 L'auteur est méthodologiste principal à la Division des méthodes d'enquêtes-entreprises, Statistique Canada.

b) Lorsque Y est une variable de grandeur, l'*information confidentielle* d'un particulier i sur Y à l'intérieur d'une cellule de **B** incluant i est la loi de répartition des proportions des contributions des unités de cette cellule dans **B**.

2.3 Remarque. L'information confidentielle est donc une loi de probabilité. Si par exemple, **B** est un tableau à quatre dimensions dont une est la variable d'intérêt Y, le conditionnement se fait en croisant les trois autres dimensions. Ce conditionnement représente l'information connue apriori par un "divulgueur". Nous prenons donc une position très conservatrice en agissant de la sorte. Une extension de la définition est possible pour les produits en forme de micro-données. En fait, il s'agit de limiter la tribu du conditionnement en limitant le nombre de variables la définissant. Nous pouvons voir cela comme une imposition d'une limite à la connaissance des divulgueurs potentiels sur les particuliers, ce qui est naturel.

2.4 Définition. Nous dirons que l'information confidentielle d'un particulier sur une variable Y est *à risque élevé* si la loi de probabilité de Y conditionnée par la tribu pour ce particulier est presque dégénérée. D'autre part, elle sera dite *à risque nul* ou *négligeable* si cette même probabilité est proche de la loi de probabilité uniforme.

2.5. Lorsque qu'un Bureau de statistique diffuse de l'information, il doit contrôler le "niveau" du risque généré par l'information confidentielle. Si le niveau est acceptable alors l'information peut être diffusée sans problème, sinon le Bureau doit modifier l'information pour que le risque généré devienne acceptable. Ces modifications de l'information seront appelées "règles de confidentialité". Elles doivent changer l'information confidentielle sans trop compromettre la valeur des analyses. Formellement, la définition suivante est justifiée.

2.6 Définition. Une *règle de confidentialité* est une fonction aléatoire définie sur les lois de probabilité (information confidentielle) et à valeurs dans le produit croisé de l'ensemble {0, 1} par l'ensemble des lois de probabilité.

2.7 Remarque. L'interprétation du singleton {0} est la suppression de l'information. L'interprétation de {1} est la publication de l'information (modifiée ou non). Lorsque les marginales sont incluses dans un tableau, la suppression de cellules n'est pas équivalente à la suppression de l'information. En fait, la suppression de cellules dans ce cas équivaut à publier des intervalles de valeurs possibles. Les règles de confidentialité doivent s'assurer que ces intervalles soient assez grands.

2.8. Le but d'une règle de confidentialité est donc de modifier l'information confidentielle pour en réduire le niveau de risque. Parler de "niveau de risque" revient à définir un ordre sur les règles de confidentialité. Nous pourrions définir formellement le risque de divulgation comme un ordre linéaire sur les règles de confidentialité ou tout simplement

2.9 Définition. Un *risque de divulgation* est une fonction réelle définie sur les images des règles de confidentialité ayant {1} comme première composante.

2.10. Il existe une multitude de risques de divulgation. Par exemple, le maximum des composantes d'une loi de probabilité discrète en est un. Serait-il possible objectivement de déterminer un risque de divulgation accepté de tous? Quelles en seraient ses propriétés? Premièrement, il serait souhaitable d'avoir la continuité et la symétrie par rapport aux composantes des lois de probabilité. De plus, la définition 2.4 suggère que l'image du risque soit un intervalle réel borné ayant l'image des lois uniforme et dégénérées aux extrémités de l'intervalle. Cette discussion justifie

2.11 Théorème. (Faddeev) *Soit* $(H_n)_{n\geq 1}$ *une suite de fonctions réelles où* H_n *est définie sur les lois de probabilité discrètes à n états. Supposons qu'elles satisfont les contraintes suivantes:*

1. H_n *est non négative, continue et symétrique par rapport aux composantes;*

2. $H_{n-1}(p_1, \ldots, p_n, 0) = H_n(p_1, \ldots, p_n)$

3. $H_n(p_1, \ldots, p_n) \leq H_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$

4. $H_{nm}(\pi_1, \ldots, \pi_{nm}) = H_n(p_1, \ldots, p_n) + \sum_{k=1}^{n} p_k H_m\left(\frac{\pi_{k:}}{p_k}, \ldots, \frac{\pi_{km}}{p_k}\right)$

où $p_k := \sum_i \pi_{ki}$ *et* $\sum_k p_k = 1$,

alors

$$(1) \qquad H_n(p_1, \ldots, p_n) = -\lambda \sum_{k=1}^{n} p_k \log(p_k),$$

où λ *est une quantité positive.*

Le symbole ':=' représente une affectation et non une égalité.

Preuve du théorème 2.11. Une démonstration se trouve dans [3]. Pour les lecteurs désireux d'obtenir la source, la première mention de ces axiomes pour caractériser (1) se trouve dans [1]. ///

2.12 Remarque. En utilisant la propriété (4) du théorème 2.11 nous pouvons déterminer, dans le cas des tableaux de grandeurs, le risque de sommes de cellules (agrégations faites de parties disjointes) en sachant seulement le risque de ses parties. C'est une propriété qui rendrait la détermination de patrons de suppressions plus efficiente.

2.13. Posons $\lambda := 1$ dans la formule (1). L'image de la formule résultante est l'intervalle [0, log(n)]. La valeur 0 n'est prise que pour les lois de probabilité dégénérées et la valeur maximale (log(n)) seulement par la loi de probabilité uniforme. Nous aimerions cependant que l'image soit indépendante de n pour que nous puissions être en mesure de comparer le risque pour plusieurs règles de confidentialité. Nous définissons donc le "risque de divulgation entropique" par

2.10. Il existe une multitude de risques de divulgation. Par exemple, le maximum des composantes d'une loi de probabilité discrète en est un. Serait-il possible objectivement de déterminer un risque de divulgation accepté de tous? Quelles en seraient ses propriétés? Premièrement, il serait souhaitable d'avoir la continuité et la symétrie par rapport aux composantes des lois de probabilité. De plus, la définition 2.4 suggère que l'image du risque soit un intervalle réel borné ayant l'image des lois uniforme et dégénérées aux extrémités de l'intervalle. Cette discussion justifie

2.11 Théorème. (Faddeev) *Soit* $(H_n)_{n\geq 1}$ *une suite de fonctions réelles où* H_n *est définie sur les lois de probabilité discrètes à n états. Supposons qu'elles satisfont les contraintes suivantes:*

1. H_n *est non négative, continue et symétrique par rapport aux composantes;*
2. $H_{n+1}(p_1, \ldots, p_n, 0) = H_n(p_1, \ldots, p_n)$
3. $H_n(p_1, \ldots, p_n) \leq H_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$
4. $H_{nm}(\pi_1, \ldots, \pi_{nm}) = H_n(p_1, \ldots, p_n) + \sum_{k=1}^{n} p_k H_m\left(\frac{\pi_{k1}}{p_k}, \ldots, \frac{\pi_{km}}{p_k}\right)$

où $p_k := \sum_i \pi_{ki}$ [2] *et* $\sum_k p_k = 1$,

alors

$$(1) \qquad H_n(p_1, \ldots, p_n) = -\lambda \sum_{k=1}^{n} p_k \log(p_k),$$

où λ *est une quantité positive.*

Preuve. Une démonstration se trouve dans [3]. Pour les lecteurs désireux d'obtenir la source, la première mention de ces axiomes pour caractériser (1) se trouve dans [1]. ///

2.12 Remarque. En utilisant la propriété (4) du théorème 2.11 nous pouvons déterminer, dans le cas des tableaux de grandeurs, le risque de sommes de cellules (agrégations faites de parties disjointes) en sachant seulement le risque de ses parties. C'est une propriété qui rendrait la détermination de patrons de suppressions plus efficiente.

2.13. Posons $\lambda := 1$ dans la formule (1). L'image de la formule résultante est l'intervalle [0, log(n)]. La valeur 0 n'est prise que pour les lois de probabilité dégénérées et la valeur maximale (log(n)) seulement par la loi de probabilité uniforme. Nous aimerions cependant que l'image soit indépendante de n pour que nous puissions être en mesure de comparer le risque pour plusieurs règles de confidentialité. Nous définissons donc le "risque de divulgation entropique" par

---

2 := représente une affectation et non une égalité.

2.14 Définition. Le *risque de divulgation entropique* ou *risque entropique* est le risque donné par la formule

$$D_n(p_1,\ldots,p_n) := 1 - \frac{1}{\log(n)} \sum_{k=1}^{n} p_k \log(p_k).$$

2.15. La section suivante fera la critique d'une règle de confidentialité pour des tableaux de grandeurs en utilisant le risque de divulgation entropique.

3 Règle de confidentialité basée seulement sur le nombre de contributeurs

3.1. Il est intuitivement accepté que le risque de divulgation devrait diminuer lorsque le nombre de contributeurs n (dans une cellule) augmente. Est-ce vrai? Cette section se propose de répondre à cette question. Pour ce faire, nous allons regarder le comportement de l'espérance mathématique du risque entropique lorsque nous faisons varier le nombre de contributeurs. Supposons que les produits de diffusion étudiés ici sont des tableaux sur de petites aires géographiques donnant les revenus des particuliers pour des sous-groupes de la population. L'information confidentielle d'une cellule (qui est la même pour tous les contributeurs de la cellule) est la loi de répartition du revenu entre les contributeurs. La règle de confidentialité est pour la deuxième composante la fonction identité tandis que la première composante vaut 0 si le nombre de contributeurs est inférieur ou égal à n et vaut 1 autrement.

3.2. Nous supposons que la grandeur publiée d'une cellule ayant n contributeurs est la réalisation de la somme de n variables aléatoires positives indépendantes et identiquement distribuées. Commençons par le théorème suivant.

3.3 Théorème. *Soit $X_1,\ldots,X_n$ des variables aléatoires positives indépendantes et identiquement distribuées.*
Soit

$$D_n(X_1,\ldots,X_n) := \frac{\log(n) + \sum_{i=1}^{n} \frac{X_i}{\sum_j X_j} \log\left(\frac{X_i}{\sum_j X_j}\right)}{\log(n)},$$

alors

$$E[D_n] \le \frac{\log\left(n^2 Var\left(\frac{X_1}{X_1+\ldots+X_n}\right)+1\right)}{\log(n)}.$$

Preuve. Soit $P_i := \frac{x_i}{\sum_j x_j}$. Ces P_i sont des variables aléatoires identiquement distribuées.

Elles sont toutefois dépendantes car $\sum_j P_j = 1$.

$E[D_n]$ existe car

$$E[|P_i \log(P_i)|] \leq \frac{1}{e} < \infty.$$

L'espérance mathématique de D_n s'écrit en termes de P_i de la façon suivante

$$E[D_n] = 1 + \frac{1}{\log(n)} E\left[\sum_j P_j \log(P_j)\right].$$

En utilisant une inégalité bien connue entre les moyennes arithmétique et géométrique

$$\sum_{i=1}^{n} q_i x_i \geq \prod_{i=1}^{n} x_i^{q_i} \qquad x_i, q_i \geq 0, \sum_j q_j = 1,$$

nous obtenons pour $q_i = x_i = p_i$ la relation suivante

$$\sum_{i=1}^{n} P_i^2 \geq \prod_{i=1}^{n} P_i^{P_i}.$$

En appliquant le logarithme, nous obtenons:

$$\log\left(\sum_{i=1}^{n} P_i^2\right) \geq \sum_{i=1}^{n} P_i \log(P_i).$$

Ce qui donne

$$E[D_n] \leq 1 + \frac{1}{\log(n)} E\left[\log\left(\sum_{i=1}^{n} P_i^2\right)\right].$$

Puisque le logarithme est concave, et que le $\log(\sum_j P_j^2)$ est intégrable, nous obtenons par l'inégalité de Jensen inversée que

$$E[D_n] \leq 1 + \frac{1}{\log(n)} \log\left(\sum_{i=1}^{n} E[P_i^2]\right)$$

$$= 1 + \frac{1}{\log(n)} \log(nE[P_i^2]).$$

Puisque $E[P_1] = 1/n$, nous obtenons ainsi

$$E[D_n] \leq 1 + \frac{\log(n^2 Var(P_1) + 1) - \log n}{\log(n)},$$

ce qu'il fallait démontrer. ///

3.4 Remarque. Une condition suffisante pour que l'espérance mathématique du risque entropique tende vers 0 est de demander que la variance des proportions soit de l'ordre de $1/n^2$, c'est-à-dire

$$n^2 Var(P_1) < constante$$

lorsque n tend vers l'infini.

3.5. Prenons un exemple instructif. Supposons que les variables aléatoires $X_1,\ldots,X_n$ suivent une loi exponentielle de paramètre λ. Sous cette hypothèse, on peut montrer aisément que la loi de la proportion P_1 a la densité suivante

$$f_P(p) := (n-1)(1-p)^{(n-2)} \qquad 0 < p < 1.$$

La variance de P_1 étant $\frac{n-1}{n^2(n+1)}$, nous avons que $E[D_n] \to 0$. Ce qui est surprenant cependant, c'est que la décroissance de l'espérance mathématique n'est que de l'ordre de 1/log(n). Pour voir cela, calculons au complet l'espérance mathématique du risque entropique.

$$E[D_n] = 1 + \frac{1}{\log(n)} E\left[\sum_i P_i \log(P_i)\right]$$

$$= 1 + \frac{n}{\log(n)} E[P_1 \log(P_1)]$$

$$= 1 + \frac{n}{\log(n)} (n-1) \int_0^1 p(1-p)^{n-2} \log(p)\, dp$$

$$= 1 + \frac{n}{\log(n)} \left[\frac{1}{n} + \int_0^1 (1-p)^{n-1} \log(p)\, dp\right].$$

En brisant le polynôme, nous obtenons

$$E[D_n] = 1 + \frac{n}{\log(n)} \left[\frac{1}{n} + \sum_{j=0}^{n-1} (-1)^j \binom{n-1}{j} \int_0^1 p^j \log(p)\, dp\right]$$

$$= 1 + \frac{n}{\log(n)} \left[\frac{1}{n} + \sum_{j=0}^{n-1} (-1)^{j+1} \binom{n-1}{j} \frac{1}{(j+1)^2}\right]$$

$$= 1 + \frac{1}{\log(n)} \left[1 - \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} \frac{1}{k}\right].$$

En utilisant l'identité [4, p. 141]

$$\sum_{k=1}^{n} \frac{(-1)^{k+1} \binom{n}{k}}{k} = \sum_{k=1}^{n} \frac{1}{k},$$

nous obtenons

$$E[D_n] = \frac{1 + \log(n) - \sum_{k=1}^{n} \frac{1}{k}}{\log(n)}.$$

Ce qui donne

$$\lim_{n\to\infty} \log(n) E[D_n] = 1 - \gamma,$$

où γ est la constante d'Euler.

3.6. Tout le monde s'accorde pour affirmer que la loi exponentielle a une trop petite queue pour simuler correctement les réalisations des proportions des revenus des particuliers. Ce qui revient à dire que l'espérance mathématique réelle du risque entropique, si elle tend vers 0, converge plus lentement que 1/log(n).
Les réalisations des revenus des particuliers sont bien simulées par la loi Pareto, laquelle possède la densité suivante

$$f_X(x) := \nu a^{\nu} x^{-(\nu+1)} \qquad a > 0, 1 < \nu < 2.$$

L'hypothèse $1 < \nu < 2$ découle de la propriété de stabilité observée par les réalisations du revenu [2]. Nous prendrons tout de même $\nu = 1$ dans ce qui suit car c'est la densité qui possède la plus grande queue, donc les risques plus élevés. On verra par la suite que le comportement de l'espérance mathématique, par voie de simulations, est similaire à un facteur près.

3.7 Lemme. *Soit $X_1, \ldots, X_n$ des variables aléatoires i.i.d. dont la densité est donnée par*

$$f_X(x) := a x^{-2} \qquad x > a > 0.$$

Soit $P_1 := \frac{X_1}{X_1 + \ldots + X_n}$

alors la loi régissant P_1 possède la densité suivante sur l'intervalle (1/n,1]

$$f_{P_1}(p_1) := c_n p_1^{-2} \qquad p_1 \in \left(\frac{1}{n}, 1\right],$$

où

$$c_n := \int_1^{\infty} \cdots \int_1^{\infty} \frac{1}{(y_2 \cdot \ldots \cdot y_n)^2} \frac{1}{(y_2 + \ldots + y_n)} dy_2 \ldots dy_n.$$

Preuve. La loi jointe de $X_1,\ldots,X_n$ est donnée par la densité

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) := a^n \frac{1}{(x_1 \cdot \ldots \cdot x_n)^2} \qquad x_i > a, i = 1, \ldots, n.$$

Faisons formellement un changement de variables. Nous verrons par la suite sa justification. Posons $P_1 := \frac{X_1}{X_1 + \ldots + X_n}$ et $P_i := X_i$ pour $i > 1$. La transformation inverse est donnée par $X_1 = \frac{P_1}{(1 - P_1)}(P_2 + \ldots + P_n)$ et $X_i = P_i$ pour $i>1$, et le jacobien de la transformation est donné par

$$J := \frac{(P_2 + \ldots + P_n)}{(1 - P_1)^2},$$

et les bornes d'intégration sont devenues $0 < P_1 < 1$ et $(P_2 + \ldots + P_n) > \frac{(1-P_1)}{P_1}a$, $P_i > a$.

Donc formellement nous obtenons la formule

$$f_{P_1}(p_1) = \frac{a^n}{p_1^2} \int_{\substack{P_2+\ldots+P_n > \frac{(1-p_1)}{p_1}a \\ P_i > a}} \frac{1}{(P_2 \cdot \ldots \cdot P_n)^2} \frac{1}{(P_2 + \ldots + P_n)} dP_2 \ldots dP_n.$$

Sur l'intervalle (1/n, 1], nous obtenons

$$f_{P_1}(p_1) = \frac{a^n}{p_1^2} \int_a^\infty \ldots \int_a^\infty \frac{1}{(P_2 \cdot \ldots \cdot P_n)^2} \frac{1}{(P_2 + \ldots + P_n)} dP_2 \ldots dP_n.$$

En transformant de nouveau par $Y_i := \frac{P_i}{a}$ pour $i > 1$, nous obtenons la définition de c_n. Le changement de variables sur l'intervalle (1/n,1] est justifié car $f_{P_1}(\cdot)$ est intégrable sur (1/n,1]. Ce qui complète la démonstration. ///

3.8. On peut voir aisément que $c_n < 1/n$. Il est également possible de trouver une borne inférieure efficace. En fait

$$\frac{c_n}{\left(1 - \frac{1}{n}\right)^{n-1}} = \frac{1}{\left(1 - \frac{1}{n}\right)^{n-1}} \int_1^\infty \ldots \int_1^\infty \frac{1}{(x_2 \cdot \ldots \cdot x_n)^2} \frac{1}{(x_2 + \ldots + x_n)} dx_2 \ldots dx_n$$

$$\geq \frac{1}{\left(1 - \frac{1}{n}\right)^{n-1}} \int_1^n \ldots \int_1^n \frac{1}{(x_2 \cdot \ldots \cdot x_n)^2} \frac{1}{(x_2 + \ldots + x_n)} dx_2 \ldots dx_n$$

$$\geq \frac{1}{\frac{1}{\left(1 - \frac{1}{n}\right)^{n-1}} \int_1^n \ldots \int_1^n \frac{1}{(x_2 \cdot \ldots \cdot x_n)^2} (x_2 + \ldots + x_n) dx_2 \ldots dx_n}$$

$$= \frac{1}{\left(1 - \frac{1}{n}\right)^{-1} (n-1) \log(n)}.$$

Nous avons utilisé l'inégalité de Jensen dans le développement précédent. Ainsi nous avons obtenu la borne inférieure

$$c_n \geq \frac{\left(1 - \frac{1}{n}\right)^n}{(n-1)\log(n)}.$$

En utilisant cette borne, nous pouvons obtenir une borne inférieure pour la variance de P_1. En fait,

$$Var[P_1] = E[P_1^2] - \frac{1}{n^2}$$

$$\geq \int_{\frac{1}{n}}^{1} p_1^2 f_{P_1}(p_1) dp_1 - \frac{1}{n^2}$$

$$= c_n \left(1 - \frac{1}{n}\right) - \frac{1}{n^2}$$

$$\geq \frac{\left(1 - \frac{1}{n}\right)^{n+1}}{(n-1)\log(n)} - \frac{1}{n^2}.$$

Nous voyons que Var[P_1] n'est pas de l'ordre de n^{-2}. Nous allons montrer que l'espérance mathématique du risque entropique est bornée inférieurement par une constante positive relativement élevée.

3.9 Théorème. *Sous les hypothèses du lemme 3.7, nous avons*

$$\liminf_{n \to \infty} E[D_n] \geq cte > 0.$$

Preuve. Supposons le contraire, alors il existe une sous-suite (n_k) telle que

$$\lim_{n_k \to \infty} E[D_{n_k}] = 0.$$

Posons

$$g_n(p) := -\frac{n}{\log(n)} p \log(p).$$

L'hypothèse implique que

$$\lim_{n_k \to \infty} E[g_{n_k}] = 1,$$

où l'espérance mathématique est calculée en se servant de la loi P_1. Puisque $g_n \leq 1$ et que $g_n = 1$ seulement au point 1/n, cela implique que toute la masse de probabilité va se greffer au point 0 lorsque n_k tend vers l'infini. Cela n'est toutefois pas possible car

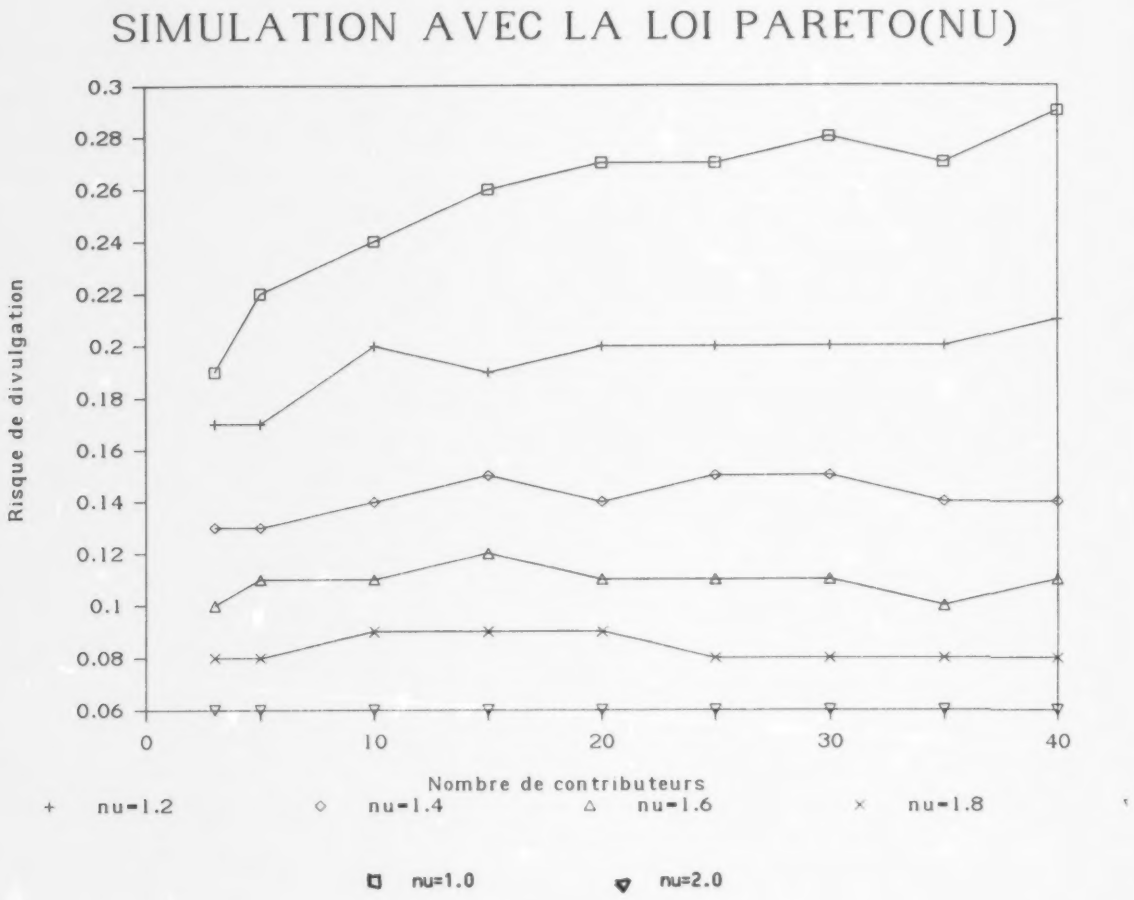
$$\int_{\frac{1}{n_k}}^{1} g_{n_k}(p) f_{P_1}(p) dp = -\frac{c_{n_k} n_k}{\log(n_k)} \int_{\frac{1}{n_k}}^{1} \frac{\log(p)}{p} dp$$

$$= \frac{1}{2} \frac{c_{n_k} n_k}{\log(n_k)} \log^2(n_k)$$

$$\geq \frac{1}{2} \left(1 - \frac{1}{n_k}\right)^{n_k} \frac{n_k}{n_k - 1} > \frac{e^{-1}}{3.}$$

En rejetant l'hypothèse, nous obtenons le résultat. ///

3.10. Pour donner une idée de la constante, nous avons simulé des réalisations du risque entropique pour différents n, de même en faisant aussi varier le paramètre ν de la loi de Pareto. Le graphique suivant donne les résultats obtenus de la simulation.



Chaque point du graphique a été obtenu par 1000 simulations de risques entropiques auxquels nous avons extrait la moyenne. Nous pouvons voir que l'espérance mathématique du risque entropique ne dépend presque pas du nombre de contributeurs. Pour ν proche de 1, cela peut représenter un danger réelle si d'autres mesures de confidentialité ne sont pas envisagées.

3.11. Nous avons donc montré dans cette section que la règle de confidentialité qui supprime les cellules ayant un nombre de contributeurs trop faible n'est pas satisfaisante. De plus, nous montrons clairement que le danger est relativement stable lorsque nous faisons varier le nombre de contributeurs, ce qui va à l'encontre de l'idée intuitive du risque de divulgation véhiculée présentement.

Références

[1] **FADDEEV D. K.**, (1957), Zum Begriff der Entropie eines endliche n Wahrscheinlichkeitsschemas. *In Arbeiten zur Informationstheorie*, Vol. 1, pp. 86-90, VEB, Berlin

[2] **MANDELBROT B.**, (1961), Stable Paretian Random Functions and the Multiplicative Variation of Income. *Econometrica*, Vol. 29, No. 4, pp. 517-543

[3] **GUIASU S.**, (1977), Information Theory with Applications. *Advance Book Program*, McGraw-Hill, London, p. 439

[4] **MERCIER A.**, (1984), Quelques identités de l'analyse combinatoire. *Descrete Mathematics*, Vol.49, North-Holland, pp. 139-149

LE TRAITEMENT STATISTIQUE POUR LA CONFIDENTIALITÉ DES MICRODONNÉES AU BSQ

Sylvie Daneault
Louise Guilbault
Bun Luk Khuong

Bureau de la statistique du Québec

INTRODUCTION

L'un des principes à la base de l'activité de tout organisme statistique est l'obligation qui lui est faite d'assurer la protection des renseignements individuels qu'il recueille dans l'exercice de ses fonctions. C'est la certitude de cette protection qui fonde la confiance du public dans l'organisme et l'amène à collaborer de bonne foi aux enquêtes de ce dernier.

De par sa loi, le Bureau de la statistique du Québec recueille et a accès à des renseignements à caractère confidentiel. Ce droit de cueillette et d'accès aux données confidentielles est contrebalancé par des règles de discrétion très rigides qui protègent les données relatives à un individu ou à une organisation. Les données transmises par les personnes, les ministères, organismes et entreprises au Bureau de la statistique du Québec sont assujetties aux règles de discrétion imposées par la Loi sur le Bureau de la statistique, qui ne permettent pas la divulgation indirecte de renseignements, encore moins la transmission de renseignements individuels ou nominatifs (L.R.Q. chapitre B-8).

Avec le recours de plus en plus large aux fichiers administratifs comme source de données et l'utilisation généralisée de l'informatique, qui facilite la manipulation de masses importantes de données ainsi que le couplage d'informations en provenance de fichiers différents, la nécessité de protéger l'intégrité des données recueillies et d'empêcher la divulgation de données confidentielles se fait encore plus sentir.

De plus, les nouvelles enquêtes du BSQ sont souvent des projets conjoints de deux partenaires ou plus. Cette tendance ne fait que compliquer encore davantage le problème de préservation de la confidentialité car tous les partenaires veulent avoir la possibilité d'accéder aux données de base (appelées couramment microdonnées) afin de faire des analyses qui répondent à leurs besoins spécifiques.

Pour faire face à ce problème, deux alternatives s'offrent au BSQ:

1° La voie légale: elle ne consiste pas à empêcher les utilisateurs d'accéder aux informations voulues. Il s'agit plutôt d'avoir le consentement du répondant pour la transmission des données requises.

2° Les techniques statistiques de masquage des données: actuellement, nous sommes au stade embryonnaire seulement. Notre recherche se limite surtout à l'évaluation des impacts de ces méthodes sur les données de base plus que sur le degré de confidentialité lié à ces méthodes; degré qui est difficile à mesurer.

LES TECHNIQUES STATISTIQUES ÉTUDIÉES AU BSQ POUR PRESERVER LA CONFIDENTIALITÉ

Un fichier de données provenant d'un échantillon plutôt que d'un recensement et ne contenant pas l'identification des unités (individus ou entreprises) le constituant, n'est pas une mesure de protection suffisante de la confidentialité. Il est possible d'identifier par unicité, par appariement ou par connaissance de la réponse, une unité appartenant à un tel fichier. Les techniques statistiques de masquage des données sont donc utilisées pour procurer un niveau de confidentialité acceptable dans de tels fichiers, sans toutefois trop distortionner les données.

Deux techniques bien connues ont été étudiées jusqu'ici par le BSQ; l'échange de données (ou permutation) et l'arrondissement aléatoire, qui sont toutes deux faciles d'application et à faibles coûts d'implantation. Bien que l'arrondissement soit surtout connu comme étant une façon de préserver la confidentialité des macrodonnées (tableaux de fréquences), l'approche qui nous intéresse ici s'effectue sur des microdonnées. La première technique s'avère surtout efficace pour des données de type catégorique, tandis que la seconde s'applique à des données de type continu.

1. Échange de données

L'échange de données consiste à intervertir les données pour différents répondants, sur un ensemble de variables "clés" sélectionnées qui sont plus révélatrices que les autres.

1.1 Échange aléatoire

Cet échange peut se faire de façon aléatoire pour l'ensemble des répondants ou de façon aléatoire pour différents répondants qui ont des caractéristiques identiques. Par exemple: même localité géographique, secteur d'activité économique, strate....

1.1.1 Exemples

Échange de façon aléatoire:

VAR (avant masquage)	VÂR (après masquage)	STRATE
3	3	1
7 ↰	10	1
10 ↲	7	2

Échange de façon aléatoire dans la strate:

VAR (avant masquage)	VÂR (après masquage)	STRATE
3 ↰	7	1
7 ↲	3	1
10	10	2

1.2 Échange par variable et par groupe de variables

L'échange de données peut s'effectuer en permutant les données indépendamment pour chaque variable, ou en permutant les données pour une partie des variables de l'enregistrement. La première méthode s'avère un peu plus confidentielle que la seconde qui elle, est sensiblement plus facile à appliquer.

1.2.1 Exemples

Échange par variable:

VAR1 (avant masquage)	VAR2	VÂR1 (après masquage)	VÂR2
3 ↰	100	7	100
7 ↲	200 ↰	3	3 000
10	3 000 ↲	10	200

Échange par groupe de variables:

VAR1 (avant masquage)	VAR2	VÂR1 (après masquage)	VÂR2
3 ↰	100 ↰	7	200
7 ↲	200 ↲	3	100
10	3 000	10	3 000

2. Arrondissement

L'arrondissement consiste à arrondir chaque donnée par un multiple d'un entier proportionnel au degré de confidentialité de la variable, cet entier représente la base d'arrondissement. Il existe deux types d'arrondissement: l'arrondissement systématique (ou classique) et l'arrondissement aléatoire qui est une variante de l'arrondissement systématique.

2.1 Arrondissement aléatoire vs arrondissement systématique

L'arrondissement systématique se fait de façon à minimiser l'écart entre la valeur masquée et la valeur originale, soit minimiser l'information perdue. Cette méthode risque cependant d'introduire un biais lors de l'estimation. De plus, les règles strictes de l'arrondissement systématique pour minimiser l'écart peuvent servir à déduire les valeurs arrondies au préalable.

Le but de l'arrondissement aléatoire est de rendre improbable et indéterminable l'exactitude de chaque donnée par l'introduction d'une erreur non biaisée d'une petitesse raisonnable. Cette méthode consiste donc à arrondir les données quelques fois vers le haut et quelques fois vers le bas, de sorte qu'à long terme les erreurs négatives et positives s'équilibrent pour donner un biais nul. Chaque décision d'arrondir vers le haut ou vers le bas doit être aléatoire.

Le BSQ s'est donc tourné vers l'étude de l'arrondissement aléatoire, cette méthode correspondant plus au besoin du bureau en matière de confidentialité pour les raisons suivantes: facilité d'application, non biaisée pour l'ensemble des répondants lors de l'estimation et surtout, plus confidentielle.

2.2 Formule théorique de l'arrondissement aléatoire

On suppose que toutes les données d'une variable sont arrondies à des multiples d'une valeur entière quelconque, appelée base d'arrondissement. Ainsi, pour chaque donnée, on a:

Si NBALEA < RESTE (=VAR modulo BASE)
Alors VÂR = VAR + BASE - RESTE (arrondie vers le haut)
Sinon VÂR = VAR - RESTE (arrondie vers le bas)

où VAR: variable originale à masquer
VÂR: variable masquée
BASE: base d'arrondissement
NBALEA: nombre aléatoire compris entre 0 et BASE
RESTE: reste à arrondir vers le haut ou vers le bas

En pratique pour un utilisateur de SAS, les nombres aléatoires peuvent être générés par la fonction RANUNI. Un argument égal à zéro dans la fonction produit une valeur initiale qui est déterminée par l'heure exacte de départ du programme d'ordinateur correspondant. Par la suite, la fonction retourne une observation générée par le nombre aléatoire précédent. Ainsi, les nombres aléatoires sont bel et bien imprévisibles. RANUNI engendre des valeurs comprises entre 0 et 1. La multiplication de ces nombres par la base d'arrondissement nous donne alors des nombres aléatoires compris entre 0 et BASE.

2.2.1 Exemples

Exemple 1 - Arrondissement aléatoire

Données pour la variable nombre d'employés : 3, 7, 10
Nombres aléatoires générés compris entre 0 et 1: 0,2; 0,7; 0,1
Base d'arrondissement choisie en fonction de l'étendue des données: 3

1) Arrondissement de la première valeur de la variable nombre d'employés

NBALEA = 0,2x3 < RESTE = 3 modulo 3
= 0,6 < = 0
FAUX

Alors, VÂR = 3 - 0 = 3 (arrondie vers le bas, mais la valeur reste intacte)

2) Arrondissement de la deuxième valeur

NBALEA = 0,7x3 < RESTE = 7 modulo 3
= 2,1 < = 1
FAUX

Alors, VÂR = 7 - 1 = 6 (arrondie vers le bas)

3) Arrondissement de la troisième valeur

NBALEA = 0,1x3 < RESTE = 10 modulo 3
= 0,3 < = 1
VRAI

Alors, VÂR = 10 + 3 - 1 = 12 (arrondie vers le haut)

Résultats :

VAR (avant masquage)	VÂR (après masquage)
3	3
7	6
10	12

Exemple 2 - Arrondissement systématique

VAR (avant masquage)	VÂR (après masquage)
3	3
7	6
10	9

2.3 Arrondissement aléatoire simple

L'arrondissement aléatoire telle que décrit précédemment ne possède qu'une seule base d'arrondissement, c'est l'arrondissement aléatoire simple. Le problème des données extrêmes rend difficile le choix d'une bonne base d'arrondissement. Il faut trouver une base qui nous permette un juste compromis entre une protection suffisante et une distortion minimale des données. La base étant différente pour chaque variable à masquer, à cause de l'étendue différente des variables, le choix de la base optimale est toujours à recommencer.

2.4 Arrondissement aléatoire à bases multiples

Pour essayer de remédier au problème de confidentialité soulevé par l'arrondissement aléatoire simple, l'arrondissement aléatoire à bases multiples a été étudié. Cette méthode d'arrondissement consiste à arrondir toutes les données de la variable à masquer selon la base 3 et à conserver les données masquées qui respectent le critère minimal de confidentialité suivant: au moins 3 observations masquées doivent posséder la même valeur arrondie à l'intérieur d'une même strate (niveau le plus fin de désagrégation). Un regroupement de strates est effectué au préalable si la strate contient moins

de 3 observations non-manquantes. Pour toutes les données qui ne respectent pas le critère minimal de confidentialité, il faut refaire l'arrondissement sur la valeur originale non-masquée, mais selon une base d'arrondissement égale à la base précédente plus 1 (ici, base=4). Parmi les nouvelles observations masquées, on conserve celles qui respectent le critère minimal de confidentialité et on refait l'arrondissement à une nouvelle base incrémentée de 1 par rapport à la précédente pour les valeurs originales des données qui ne respectent pas le critère minimal et ainsi de suite, jusqu'à une base jugée optimale. Après l'application de cette base optimale, il ne reste que des données dispersées dans le fichier, pour la plupart des données extrêmes.

L'arrondissement aléatoire à bases multiples pourrait se faire avec une base maximale plutôt qu'optimale. Une base maximale permettrait à toutes les données d'être arrondies, sauf celles qui ne peuvent respecter le critère minimal de confidentialité (lorsqu'il reste moins de 3 données non-arrondies). Il faudrait alors pour ces données faire quand même de l'imputation à la valeur arrondie la plus près dans la strate. Une base maximale viendrait augmenter les coûts d'exploitation et ne permettrait pas un niveau de confidentialité plus élevé.

L'arrondissement aléatoire à bases multiples comporte un léger biais dans l'estimation pour l'ensemble des répondants, contrairement à l'arrondissement aléatoire simple. Seulement une partie des valeurs arrondies sont conservées à chaque "tour" d'arrondissement et ainsi, il n'est pas certain que les erreurs positives et négatives s'équilibrent pour donner un biais nul.

2.5 Imputation en fonction de la strate

L'imputation vient compléter l'arrondissement aléatoire à bases multiples comme technique de masquage. Les données restantes non-arrondies sont imputées à la valeur la plus près parmi les valeurs déjà arrondies dans la strate et, si aucune valeur n'est arrondie dans la strate, alors il faut imputer ces valeurs à zéro ou à la valeur moyenne dans la strate. Ce dernier recours est justifié par le fait que l'imputation à la valeur moyenne vient ajouter une erreur soit positive ou négative, contrairement à l'erreur toujours négative qui découle de l'imputation à zéro.

L'imputation à la valeur arrondie la plus près dans la strate respecte le critère minimal de confidentialité car l'arrondissement fait au préalable sur les données dépend de ce critère. De plus, l'imputation à zéro ou à la valeur moyenne dans les strates où aucune valeur n'a été arrondie répond aussi au critère; la restriction sur le nombre minimal d'observations par strate (3) en est la preuve.

Il existe une lacune à cette méthode; les données qui sont imputées à la valeur arrondie la plus près dans la strate et pour lesquelles cette valeur est relativement éloignée. Il en résultera un biais assez important. Pour remédier à cet inconvénient, on impute la valeur moyenne de toutes les données non-arrondies dans une même strate. Si le critère minimal de confidentialité n'est pas respecté, on doit imputer la valeur arrondie la plus près à ces données. Il est toutefois préférable, pour ne pas trop biaiser

les résultats, de séparer dans chaque strate les valeurs extrêmes de la distribution de la variable à l'étude pour ainsi calculer différentes moyennes. Par exemple, on peut prendre dans une strate donnée 5% des valeurs à la gauche de la distribution (soit les plus petites valeurs), 90% des valeurs au centre de la distribution (soit les valeurs moyennes) et 5% des valeurs à la droite de la distribution (soit les plus grandes valeurs). La moyenne de chacune des 3 sections est calculée et on impute la valeur moyenne aux données de la section correspondante. Si le critère minimal de confidentialité n'est pas respecté, il faut regrouper des sections et si cela ne suffit pas encore, on impute la valeur arrondie la plus près aux données non-arrondies dans la strate. Donc, le nombre minimum d'observations pour calculer la valeur moyenne dans chacune des sections est de 3 et le nombre maximum est égal à 5% des observations non-masquées dans les sections contenant des données extrêmes et est égal à 90% dans l'autre section.

2.6 Exemple

Arrondissement aléatoire à bases multiples

1) Base = 3

	VAR	VÂR	STRATE
respecte le critère minimal de confidentialité ←	2	3	1
	3	3	1
	4	3	1
	4	3	2
	5	3	2
	7	6	2
	10	9	2
	31	30	2

2) Base = 4

	VAR	VÂR	STRATE
respecte le critère minimal de confidentialité ←	4	4	2
	5	4	2
	7	4	2
	10	12	2
	31	32	2

Le critère minimal de confidentialité ne pourra être respecté pour les données qui restent, les valeurs non-arrondies (10, 31) seront donc imputées à la valeur 4.

3. Résultats

Les résultats de l'étude sur l'échange de données et l'arrondissement aléatoire sont montrés pour 2 strates d'un fichier de microdonnées; ce fichier contient des données sur le salaire minimum. La strate 6 est une strate hétérogène pour la variable étudiée (V35: nombre d'employés), tandis que la strate 11 est homogène. Les estimations, les coefficients de variations (CV) et la corrélation ont donc été calculés par strate.

3.1 6 SCÉNARIOS D'ARRONDISSEMENT ALÉATOIRE À BASES MULTIPLES AVEC IMPUTATION

NCIJ1: arrondissement aléatoire jusqu'à base 15
- . + imputation à la valeur arrondie la plus près dans la strate
- . sinon (aucune valeur n'est arrondie dans la strate), + imputation à zéro

NCIJ2: idem à NCIJ1 (jusqu'à base 10)

NCIJ3: arrondissement aléatoire jusqu'à base 10
- . + imputation à la valeur arrondie la plus près dans la strate
- . sinon, + imputation à la moyenne dans la strate

NCIJ4: arrondissement aléatoire jusqu'à base 10
+ imputation à la valeur moyenne des valeurs non-arrondies dans la strate

NCIJ5: idem à NCIJ4 (jusqu'à base 15)

NCIJ6: arrondissement aléatoire jusqu'à base 15
+ imputation selon les 3 moyennes des 3 sous-groupes de la distribution des valeurs non-arrondies dans la strate

3.2 RÉSULTATS SUR UNE STRATE "HOMOGÈNE" (11)

Strate où il y a 57 répondants dont l'étendue des réponses pour V35 est de [2,140]

V35 = nombre d'employés
NV35 = V35 masquée
V24 = salaire versé
NV24 = V24 masquée

Technique de masquage	Estimation	CV	Corrélation NV35 avec V24;NV24
V35 initiale (=NV35)	59 040	0,10	0,79
Permutation aléatoire par:			
- Variable	40 294	0,35	-0,09;-0,11
- Groupe de variables	41 042	0,28	-0,06;0,95

Permutation aléatoire par strate par:

- Variable	59 040	0,10	-0,31;0,01
- Groupe de variables	59 040	0,10	-0,21;0,79

Arrondissement aléatoire à base simple:

- Base 3	58 861	0,10	0,79
- Base 5	58 861	0,10	0,79

Arrondissement aléatoire à bases multiples:

- Scénario NCIJ1	55 833	0,09	0,73
NCIJ2	55 363	0,09	0,71
NCIJ3	55 363	0,09	0,71
NCIJ4	58 591	0,09	0,50
NCIJ5	59 040	0,09	0,58
NCIJ6	58 966	0,10	0,76

3.3 RÉSULTATS SUR UNE STRATE TRÈS "HÉTÉROGÈNE" (6)

Strate où il y a 26 répondants dont l'étendue des réponses pour V35 est de [12,7784]

V35 = nombre d'employés
NV35 = V35 masquée
V24 = salaire versé
NV24 = V24 masquée

Technique de masquage	Estimation	CV	Corrélation NV35 avec V24;NV24
V35 initiale (NV35)	237 510	0,51	0,996

Permutation aléatoire par:

- Variable	12 854	0,51	-0,13;-0,30
- Groupe de variables	8 336	0,30	-0,15;0,88

Permutation aléatoire par strate par:

- Variable	237 510	0,51	-0,21;-0,15
- Groupe de variables	237 510	0,51	-0,06;0,99

Arrondissement aléatoire à base simple:

- Base 3	237 266	0,51	0,996
- Base 5	237 331	0,51	0,996

Arrondissement aléatoire à bases multiples:

- Scénario NCIJ1	120 835	0,05	0
NCIJ2	121 680	0,05	0
NCIJ3	121 680	0,05	0
NCIJ4	237 543	0,06	0,08
NCIJ5	237 445	0,06	0,08
NCIJ6	237 392	0,33	0,64

3.4 PRÉSERVATION DE LA CONFIDENTIALITÉ*

Technique de masquage	Confidentialité
V35 initiale	Aucune protection
Permutation aléatoire par: - Variable - Groupe de variables	Bonne protection surtout si permutation par variable
Permutation aléatoire par strate par: - Variable - Groupe de variables	Moins bonne protection que précédemment
Arrondissement aléatoire à base simple: - Base 3 - Base 5	Faible protection surtout en base 3
Arrondissement aléatoire à bases multiples: - NCIJ1 - NCIJ2 - NCIJ3 - NCIJ4 - NCIJ5 - NCIJ6	Bonne protection due au nombre de bases et à la limite minimale de 3 observations par strate

*Note: Il s'agit ici d'une "évaluation" subjective de la confidentialité.

CONCLUSION

Le problème sera toujours le suivant: trouver un bon équilibre entre la confidentialité et la préservation des résultats. Dans le cas de la permutation aléatoire, la confidentialité des données est préservée mais les résultats eux sont perturbés. Le cas inverse se produit pour la permutation aléatoire par strate et pour l'arrondissement aléatoire simple. Quant à l'arrondissement aléatoire à bases multiples avec imputation à différentes moyennes, elle préserve la confidentialité et procure des estimations (et des CV) se rapprochant de la réalité. Toutefois, à cause de l'imputation massive dans les strates hétérogènes, cette méthode détruit parfois la corrélation entre les variables.

L'étude sur l'arrondissement aléatoire à bases multiples devrait donc se poursuivre pour rendre cette méthode plus performante. Ainsi, ne serait-il pas préférable d'imputer la valeur arrondie la plus près dans les strates homogènes et les 3 valeurs moyennes dans les strates hétérogènes?

Le degré de confidentialité demeure toutefois difficile à quantifier car il dépend directement des risques d'identification par unicité, par appariement et par connaissance de la réponse, donc des risques de divulgation. Ces risques dépendent entre autres, de la taille d'échantillon et du nombre de variables diffusées. Jusqu'à présent, des études ont été faites par certains organismes pour calculer le critère de confidentialité d'une variable et pour déterminer les enregistrements à haut risque de révélation (à l'aide de la mesure d'entropie). Dans les années subséquentes, le BSQ aura donc comme mandat d'étudier plus en profondeur cet aspect de mesure de la confidentialité pour ainsi pouvoir contrer le problème de l'identification et de la divulgation, en utilisant des techniques de masquage appropriées.

RÉFÉRENCES

1. Causey B., "Approches to statistical disclosure", ASA Proceeding of the social statistics section, 1979, 306-309

2. Denis J., "Fichiers publics de microdonnées: le problème de l'identification et de la divulgation et quelques techniques palliatives", Division des méthodes d'enquêtes sociales, Statistique Canada , 1989

3. Greenberg B., "Designing a disclosure avoidance methodology for the 1990 Decennial Censuses", Office of the Director, U.S. Bureau of the Census, 1986.

4. Griffin R., Navarro A., Flores-Baez L., "Disclosure avoidance for the 1990 Census", U.S. Bureau of the Census, 1989

5. Nargundkar M.S., Saveland W., "Le respect du secret statistique par l'arrondissement aléatoire" Revue statistique du Canada, Février 1973

6. Navarro A., Flores-Baez L., Thompson J.H., "Results of data switching simulation", Statistical support division, Bureau of the Census, 1988

7. Spruill N., "The confidentiality and analytic usefulness of masked business microdata", Office of the Secretary of Defense

ANALYSE DE LA VALIDITÉ ET DE LA FIABILITÉ D'UN QUESTIONNAIRE D'OPINION[1]

Statistique Canada a élaboré un questionnaire d'étude de comportements de gestion. On y mesure les qualités d'un gestionnaire dans neuf différentes dimensions allant de la discipline personnelle à la négociation. Dans le but d'évaluer la validité des questions posées et leur fiabilité dans le temps, on peut alors avoir recours à des méthodes statistiques. On présente ici la méthodologie utilisée pour effectuer une telle étude. On décrit entre autres le développement de méthodes statistiques pour l'analyse de questions de préférence et on donne finalement les résultats de l'étude.

MOTS-CLÉS: comportements de gestion, questions d'opinion et de préférence, modèle linéaire, bootstrap.

1. INTRODUCTION

La Section du développement organisationnel de la Division des politiques du personnel de Statistique Canada a récemment élaboré un questionnaire destiné à analyser les comportements de gestion des cadres d'une entreprise. Ce questionnaire se veut un outil nouveau dans le domaine du développement des organisations. Une étude semblable a été faite par Pareek (1989) où il mesure la relation entre des facteurs de motivation et des éléments de culture organisationnelle.

Dans le questionnaire, on fait l'hypothèse que la culture organisationnelle se manifeste par des normes, ou hypothèses de base (Schein (1984)), et qu'elle a une influence déterminante sur le comportement des gestionnaires.

Le questionnaire a pour but premier d'aider le gestionnaire à améliorer sa contribution à l'organisation. Il indique si la "culture" de l'organisation supporte ou non les comportements du gestionnaire (Est-ce qu'il existe, par exemple, une norme dans l'organisation qui renforce le fait de rencontrer ses clients?). Il aide à identifier les forces et les faiblesses du gestionnaire. Finalement, il se veut un outil de développement personnel.

2. DESCRIPTION DU QUESTIONNAIRE

Afin d'atteindre les buts mentionnés dans l'introduction, le questionnaire sur les comportements de gestion mesure les neuf dimensions suivantes: la négociation, l'écoute, le partenariat, le soutien, la direction, le contrôle, l'accomplissement, la perfection et l'innovation. Les dimensions de négociation, écoute et partenariat sont principalement centrées sur le client (ou utilisateur). D'autre part, le soutien, la direction et le contrôle sont axés sur les ressources

---

1 Pierre Lavallée, Division des méthodes d'enquêtes-entreprises, Claude Gravel, Division des politiques du personnel, Statistique Canada, Ottawa, K1A 0T6

internes tandis que l'accomplissement, la perfection et l'innovation sont centrés sur le produit (ou service). On trouve la définition des dimensions à l'annexe A.

Pour chacune de ces neuf dimensions, on mesure les cinq secteurs suivants: les forces, les distorsions, les caractéristiques personnelles, les normes positives et les normes négatives. Les forces se définissent comme les comportements qui ont habituellement un impact positif sur l'organisation. Par opposition, les distorsions sont les comportements qui ont habituellement une influence négative sur l'organisation. Par caractéristiques personnelles, on entend les traits de la personne (par exemple, être sensible aux autres versus être organisateur) qui poussent celle-ci à attribuer plus de valeurs à une façon d'agir plutôt qu'à une autre. Finalement, les normes positives se rapportent à la culture de l'organisation supportant le gestionnaire tandis que les normes négatives sont celles nuisant au gestionnaire. Ces normes sont mesurées au niveau de l'unité de travail immédiat (dénoté par T) et au niveau de la division (dénoté par D) qui constitue l'unité administrative plus large.

Le Tableau 1 représente la structure du questionnaire. On y distingue un total de 135 questions dont 126 questions d'opinions (marquées par X, T ou D) et 9 questions de préférence (chacune formée des trois choix identifiés par un trait).

Tableau 1

Secteurs	Dimensions								
	1	4	7	2	5	8	3	6	9
Forces	X X X	X X X	X X X	X X X	X X X	X X X	X X X	X X X	X X X
Distorsions	X X X	X X X	X X X	X X X	X X X	X X X	X X X	X X X	X X X
Caractéristiques personnelles	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-
	-	-	-	-	-	-	-	-	-
Normes positives	T D T D	T D T D	T D T D	T D T D	T D T D	T D T D	T D T D	T D T D	T D T D
Normes négatives	T D T D	T D T D	T D T D	T D T D	T D T D	T D T D	T D T D	T D T D	T D T D

1- Négociation 2- Écoute 3- Partenariat
4- Soutien 5- Direction 6- Contrôle
7- Accomplissement 8- Perfection 9- Innovation

Les forces, distorsions, normes positives et normes négatives sont mesurées à l'aide de questions d'opinion. A titre d'exemple, voici une question d'opinion reliée aux forces: "Je donne crédit à mes employés pour le travail bien fait. ". Le répondant inscrit le chiffre 0 si l'énoncé "ne s'applique pas", 1 s'il "s'applique peu", 2 s'il "s'applique modérément" et 3 s'il "s'applique dans une large mesure".

D'autre part, on mesure les caractéristiques personnelles par des questions de préférence. Rappelons que les caractéristiques personnelles influent sur l'ordre de priorité donnée aux dimensions par le gestionnaire. Un exemple de ces questions est le suivant: "On me dit souvent que: (a) Je travaille bien en groupe. (b) Je sais contrôler les choses. (c) Je suis innovateur. ". Le répondant marque sa préférence en utilisant les chiffres 1, 2 et 3 où 1 correspond à la réponse la moins favorisée et 3 à la plus favorisée. Dans l'exemple précédent, l'affirmation (a) est reliée à la dimension partenariat, (b) au contrôle et (c) à l'innovation.

Une fois le questionnaire rempli, on compile les résultats de chaque répondant. Pour chaque dimension, on calcule alors:

$$\text{TOTAL} = (\text{FORCES-DISTORSIONS}) \times (\text{ATTITUDES} + (\text{NORMES\_POS-NORMES\_NEG})).$$

Cette mesure permet de quantifier la performance du gestionnaire par rapport aux neuf dimensions que mesure le questionnaire. Les résultats sont finalement présentés sous forme de diagramme que l'on retrouve à l'annexe A.

3. VALIDITÉ ET FIABILITÉ

On mesure la qualité du questionnaire en termes de validité et de fiabilité. Une question est dite valide si elle mesure correctement le concept la soutenant. Par ailleurs, on dit qu'une question est fiable si sa réponse n'est pas influencée par quelque facteur externe (comme par exemple le facteur temps). Il est bon de noter ici que cette terminologie ne constitue pas un standard et qu'elle pourrait même être sujette à discussion. Elle est, quoiqu'il en soit, utilisée tout au long de cet article.

Pour mesurer la validité et la fiabilité du questionnaire, on a eu recours à un échantillon de 50 personnes choisies parmi les gestionnaires de différentes divisions de Statistique Canada. A ces 50 personnes, on a distribué un premier questionnaire (premier test) pour lequel on a obtenu 43 répondants. Deux semaines après le retour du premier questionnaire, un deuxième (deuxième test) a alors été distribué aux répondants du premier questionnaire. Ce deuxième questionnaire était en tout point conforme au premier excepté l'ordre des questions. A ce

deuxième test, 35 ont répondu. Les résultats de ces deux tests ont servi à mesurer la fiabilité du questionnaire à l'aide d'une analyse statistique décrite dans la section suivante.

Suite au deuxième test, on a distribué aux 35 répondants un troisième questionnaire (post-test) où les concepts mesurés étaient clairement expliqués dans un cours texte relié à chacun des secteurs des forces, des distorsions, des normes positives et des normes négatives pour chacune des neuf dimensions. Le répondant devait alors quantifier sur une échelle de 1 à 10 sa perception personnelle pour chacun des 45 concepts (9 dimensions × 4 secteurs). Un total de 27 répondants ont rempli le post-test.

4. ANALYSES STATISTIQUES

Pour effectuer l'analyse statistique des données recueillies auprès des répondants aux questionnaires, deux différentes techniques ont été utilisées; l'une pour les questions d'opinion et l'autre pour les questions de préférence.

4.1 ANALYSE DES QUESTIONS D'OPINION

Les questions d'opinion ont été utilisées pour mesurer les forces, les distorsions, les normes positives et les normes négatives pour chacune des neuf dimensions.

Une des difficultés reliées à l'analyse statistique de ces questions est leur nombres (126 questions) par rapport au nombre des répondants (35 pour le deuxième test et 27 pour le post-test). On remarque cependant que chaque secteur de chaque dimension comporte deux ou trois questions et que les réponses de celles-ci se mesurent dans le même sens (c'est-à-dire qu'une réponse codée 3, par exemple, pour une question ne contredit pas une autre question codée 3). Dans le but de réduire le nombre de questions d'opinion, on a donc procédé à un regroupement de questions en additionnant les résultats des deux ou trois questions de chaque secteur de chaque dimension. Ainsi, pour trois questions de réponses codées 0 à 3, on obtient un résultat codé 0 à 9. On est finalement passé de 126 à 54 résultats.

On présente ici l'analyse statistique effectuée pour la comparaison des deux premiers tests; ce qui mesure la fiabilité du questionnaire. Il faut noter que l'analyse pour la comparaison du premier test et du post-test, pour la mesure de la validité du questionnaire, est en tout point semblable.

Pour chaque secteur de chaque dimension, on a défini les variables suivantes:

Soient $x_1, \ldots, x_n$ où n=35, les résultats du premier test pour les n répondants communs aux deux tests.

Soient $y_1,\ldots,y_n$, les résultats du deuxième test.

Notons que x_i et y_i, i=1,...,n, peuvent prendre les valeurs 0 à 9.

Afin de comparer les résultats des premier et deuxième tests, on a supposé le modèle suivant:

$$y_i = \beta x_i + e_i \qquad i = 1,\ldots,n$$

où

$$e_1,\ldots,e_n \sim i.i.d.\, N(0,\sigma^2) \ .$$

On a alors vérifié en premier lieu l'hypothèse $H_0:\beta = 1$ (contre $H_1:\beta \neq 1$) au niveau significatif $\alpha = 0.05$. Si l'hypothèse ne peut être rejetée, on conclut alors que les résultats entre les deux tests sont en moyenne semblables. En deuxième lieu, on a jugé acceptable un coefficient de détermination R^2 supérieur ou égal à 75%. En effet, même si l'hypothèse précédente n'est pas rejetée, une valeur trop faible de R^2 suggère une trop grande variabilité des réponses entre les deux tests. Finalement, on a effectué une analyse des résidus visant à vérifier si les hypothèses relatives au modèle utilisé (comme par exemple la normalité) étaient valables. Pour plus de détails sur ces techniques, on peut consulter Montgomery et Peck (1982).

Les résultats de la comparaison des deux premiers tests se retrouvent à l'annexe B. Rappelons que cette comparaison vise à vérifier la fiabilité du questionnaire. On retrouve à l'annexe C les résultats de la comparaison du premier test et du post-test pour mesurer la validité du questionnaire.

A partir des résultats obtenus, on a décelé des problèmes de validité et de fiabilité au niveau du secteur des distorsions. Il semble en fait qu'il est en général difficile pour une personne de percevoir que ses comportements peuvent avoir un impact négatif. Pour solutionner ce problème, on a jugé bon de simplement éliminer les questions relatives à ce secteur.

L'analyse des résultats nous a aussi permis d'identifier des problèmes de validité et de fiabilité au niveau de la dimension de perfection. Un mélange entre les concepts de prévention et de perfection serait la cause de ces problèmes. En effet, cette dimension, appelée ici perfection, tentait en fait de mesurer des concepts de prévention. On a ainsi reformulé certaines des questions et redéfini la dimension.

Finalement, des problèmes de validité et d'autres problèmes de fiabilité ont été détectés pour des questions isolées. Ces problèmes étaient pour la plupart causés par une mauvaise formulation ou par l'usage d'expressions ambiguës. Ces questions à problèmes ont simplement été reformulées.

4.2 ANALYSE DES QUESTIONS DE PRÉFÉRENCE

Les questions de préférence ont été utilisées dans le questionnaire pour mesurer le secteur des caractéristiques personnelles, c'est-à-dire l'ordre de priorité donnée aux dimensions par le gestionnaire.

Il est à noter ici que ce secteur n'a pas été mesuré par le post-test. On n'a donc pu faire d'analyse statistique qu'au niveau des deux premiers tests. Seule la fiabilité du questionnaire est mesurée pour les questions de préférence.

Pour chaque question du secteur des caractéristiques personnelles, on a défini les variable suivantes:

Soient $\underline{x}_1, \ldots, \underline{x}_n$, les résultats du premier test où n=27 et $\underline{x}_i = (x_{i1}, x_{i2}, x_{i3})$.

Soient $\underline{y}_1, \ldots, \underline{y}_n$, les résultats du deuxième test où $\underline{y}_i = (y_{i1}, y_{i2}, y_{i3})$.

Notons que $\underline{x}_i$ et $\underline{y}_i$ peuvent prendre les valeurs suivantes: (1, 2, 3), (2, 1, 3), (3, 1, 2), (1, 3, 2), (2, 3, 1), (3, 2, 1).

Pour comparer les résultats des premier et deuxième test, on convient que l'on doit disposer d'une mesure de distance. On a défini ainsi la statistique suivante:

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$$

où $$d_i = \tfrac{1}{2}[(x_{i1}-y_{i1})^2 + (x_{i2}-y_{i2})^2 + (x_{i3}-y_{i3})^2].$$

Plus les résultats $\underline{x}_i$ et $\underline{y}_i$ sont différents, plus la valeur de d_i est grande, i=1,...,n. On note en effet que d_i peut prendre les valeurs 0, 1, 3 et 4 où $d_i = 0$ lorsque $\underline{x}_i = \underline{y}_i$ et $d_i = 4$ lorsque $\underline{x}_i$=(1, 2, 3) et $\underline{y}_i$=(3, 2, 1), et vice versa.

On a voulu vérifier l'hypothèse H_0: écart=0 (contre H_1: écart>0) au niveau significatif $\alpha = 0.05$. Le problème qui s'est posé à nous est d'obtenir la loi de $\bar{d}$ sous l'hypothèse nulle.

A partir de Efron (1982), on a pu estimer la loi de $\bar{d}$ à l'aide de la méthode du bootstrap. On a premièrement obtenu 1000 valeurs de la statistique du bootstrap

$$T^* = \sqrt{n}(\bar{D}^* - \bar{d})/S_d^*$$

où $$\bar{D}^* = \frac{1}{n} \sum_{i \in \{1,\ldots,n\}} d_i$$

et $$S_d^* = \sqrt{\frac{1}{n-1} \sum_{i \in \{1,\ldots,n\}} (d_i - \bar{D}^*)^2}.$$

La notation "$i \in \{1,\ldots,n\}$" indique que la somme s'effectue sur toutes les n unités choisies avec remise par la procédure du bootstrap. Après avoir calculé le 95ème centile des 1000 valeurs de T^*, que l'on dénote par $T^*_{0.95}$, on a rejeté H_0 lorsque

$$T = \sqrt{n}\bar{d} / S_d > T^*_{0.95}$$

où $$S_d = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (d_i - \bar{d})^2}.$$

Les résultats de ces tests statistiques se retrouvent à l'annexe D. Etant donné que dans tous les cas l'hypothèse nulle a été rejetée, on a convenu d'éliminer totalement les questions relatives à le secteur "caractéristiques personnelles".

5. CONCLUSIONS

Cet étude sur la validité et la fiabilité d'un questionnaire a permis le développement de méthodes d'analyse de questions d'opinion et de questions de préférence.

Suite à l'analyse effectuée, on a éliminé du questionnaire les secteurs des distorsions et des caractéristiques personnelles et reformulé certaines questions et concepts. Ainsi le test utilisé présentement s'avère être plus court et son utilisation jugée adéquate par les répondants.

On utilisera ce questionnaire sur les comportements de gestion tout au cours de l'année pour pouvoir présenter des résultats en psychologie industrielle.

6. BIBLIOGRAPHIE

Efron, B. (1982), "The Jackknife, the Bootstrap and Other Resampling Plans", Society for Industrial and Applied Mathematics, Philadelphie, 1982.

Pareek, U. (1989), "Motivation Analysis of Organizations - Climate (MAO-C)", The 1989 Annual: Developing Human Resources, University Associates, 1989.

Montgomery, D.C., Peck, E. (1982), "Introduction to Linear Regression Analysis", John Wiley and Sons, New York, 1982.

Schein, E.H. (1984), "Coming to a New Awareness of Organizational Culture", Sloan Management Review, hiver 1984.

ANNEXE A: Définitions et diagramme des résultats du questionnaire

COMPORTEMENTS CENTRÉS SUR LE CLIENT

NÉGOCIATION

- Clarifier les limites et les frontières
- Développer la confiance
- Être direct

ÉCOUTE

- Éliminer l'insatisfaction
- Clarifier les besoins
- Établir des contrats

PARTENARIAT

- Impliquer les clients dans le développement
- Impliquer les clients dans l'amélioration des processus
- Demander du feedback sur les façons de faire

COMPORTEMENTS CENTRÉS SUR LES RESSOURCES INTERNES

SOUTIEN

- Donner le crédit aux employés pour leur travail
- Stimuler la réflexion
- Maximiser l'utilisation des capacités

DIRECTION

- Partager sa compréhension des enjeux corporatifs
- Discuter des valeurs qui gouvernent l'organisation
- Créer la cohésion autour des buts et objectifs

CONTRÔLE

- Mesurer l'efficience
- Mesurer l'efficacité
- Mesurer l'économie

COMPORTEMENTS CENTRÉS SUR LE PRODUIT

ACCOMPLISSEMENT

- Réussir des tâches difficiles
- Obtenir ce qu'on veut en dépit de la bureaucratie
- Être systématique dans l'organisation du travail

PRÉVENTION

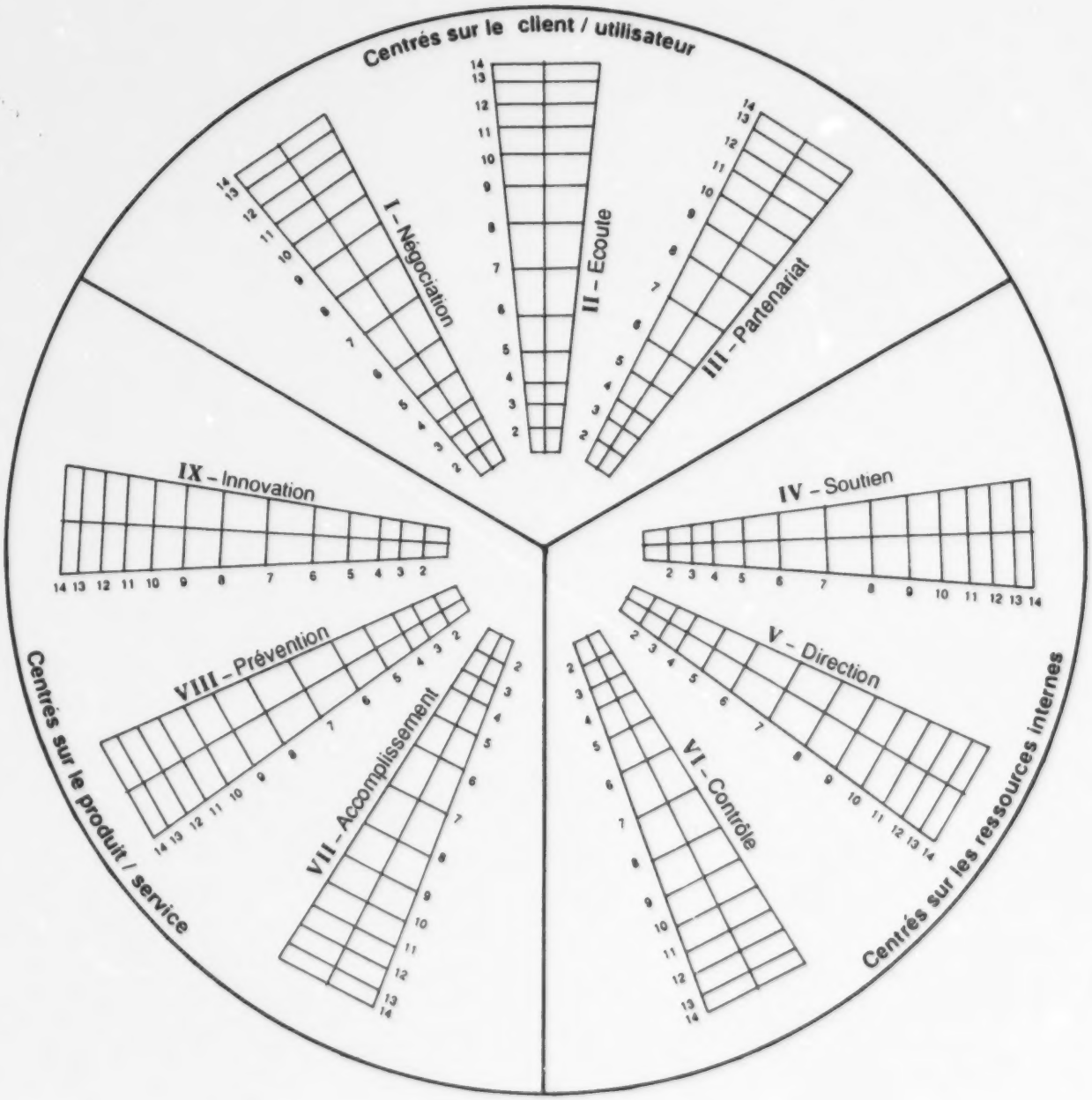
- s'assurer que ce qui provient du fournisseur rencontre les standards/politiques établis
- À chaque étape du processus, comparer le travail aux standards/politiques établis
- Établir des mesures pour l'amélioration de la qualité

INNOVATION

- Faire des choses non usuelles considérées innovatrices
- Prendre des décisions non conventionnelles
- Proposer des approches de travail non conformistes

DIAGNOSTIC DES COMPORTEMENTS DE GESTION

par
Claude Gravel



Section du développement organisationnel, STC, 1989

ANNEXE B: Résultats sur la fiabilité des questions d'opinion

(Comparaison entre les premier et deuxième tests.)

dimension	secteur	p-value	R^2	problème
1- négociation	forces	0.038	93.0	**
	distorsions	0.771	73.3	+
	normes + T	0.241	92.2	
	normes + D	0.008	84.8	**
	normes - T	0.403	84.0	
	normes - D	0.535	86.3	
2- écoute	forces	0.067	95.4	*
	distorsions	0.823	84.9	
	normes + T	0.959	96.6	
	normes + D	0.580	86.2	
	normes - T	0.652	81.7	
	normes - D	0.647	80.4	
3- partenariat	forces	0.790	90.4	
	distorsions	0.033	82.4	**
	normes + T	0.797	95.1	
	normes + D	0.249	82.9	
	normes - T	0.687	82.5	
	normes - D	0.306	80.0	
4- soutien	forces	0.958	98.4	
	distorsions	0.584	70.8	+
	normes + T	1.000	94.5	
	normes + D	0.043	82.3	**
	normes - T	0.459	68.5	+
	normes - D	0.010	58.1	** +

** significatif au niveau $\alpha = 0.05$

* significatif au niveau $\alpha = 0.10$

+ $R^2 < 75\%$

dimension	secteur	p-value	R^2	problème
5- direction	forces	0.775	93.8	
	distorsions	0.154	71.6	+
	normes + T	0.475	87.1	
	normes + D	0.715	79.1	
	normes - T	0.590	86.7	
	normes - D	0.714	81.6	
6- contrôle	forces	0.112	88.0	
	distorsions	0.002	53.2	** +
	normes + T	0.514	80.3	
	normes + D	0.208	72.5	+
	normes - T	0.619	83.7	
	normes - D	0.293	83.8	
7-accomplissement	forces	0.402	95.8	
	distorsions	0.598	89.4	
	normes + T	0.048	93.0	**
	normes + D	0.125	89.9	
	normes - T	0.285	93.2	
	normes - D	0.120	86.4	
8- perfection	forces	0.592	88.3	
	distorsions	0.001	77.1	**
	normes + T	0.176	93.0	
	normes + D	0.048	89.0	**
	normes - T	0.359	73.8	+
	normes - D	0.381	70.6	+
9- innovation	forces	0.745	63.6	+
	distorsions	0.211	84.4	
	normes + T	0.387	89.2	
	normes + D	0.554	85.3	
	normes - T	0.712	90.8	
	normes - D	0.890	91.7	

ANNEXE C: Résultats sur la validité des questions d'opinion

(Comparaison entre le premier test et le post-test.)

dimension	secteur	p-value	R^2	problème
1- négociation	forces	0.646	88.9	
	distorsions	0.329	66.1	+
	normes + T	0.183	93.5	
	normes + D	0.006	87.8	**
	normes - T	0.000	86.6	**
	normes - D	0.006	74.9	** +
2- écoute	forces	0.311	92.7	
	distorsions	0.425	80.2	
	normes + T	1.000	88.4	
	normes + D	0.841	85.4	
	normes - T	0.690	87.1	
	normes - D	0.273	76.8	
3- partenariat	forces	0.261	82.9	
	distorsions	0.034	69.1	** +
	normes + T	0.098	93.0	*
	normes + D	0.233	90.1	
	normes - T	0.245	80.1	
	normes - D	0.054	75.4	*
4- soutien	forces	0.001	96.1	**
	distorsions	0.004	70.6	** +
	normes + T	0.012	90.5	**
	normes + D	0.160	84.5	
	normes - T	0.372	43.2	+
	normes - D	0.540	63.2	+

** significatif au niveau $\alpha = 0.05$

* significatif au niveau $\alpha = 0.10$

+ $R^2 < 75\%$

dimension	secteur	p-value	R^2	problème
5- direction	forces	0.319	89.1	
	distorsions	0.094	43.9	* +
	normes + T	0.582	88.1	
	normes + D	0.748	80.6	
	normes - T	0.954	81.6	
	normes - D	0.897	83.6	
6- contrôle	forces	0.398	82.6	
	distorsions	0.302	63.5	+
	normes + T	0.038	82.6	**
	normes + D	0.160	66.3	+
	normes - T	0.622	78.4	
	normes - D	0.407	74.6	+
7-accomplissement	forces	0.765	91.7	
	distorsions	0.138	57.9	+
	normes + T	0.198	93.4	
	normes + D	0.640	84.5	
	normes - T	0.041	90.6	**
	normes - D	0.260	78.6	
8- perfection	forces	0.176	79.4	
	distorsions	0.000	65.1	**
	normes + T	0.000	95.3	**
	normes + D	0.126	93.2	**
	normes - T	0.012	70.4	** +
	normes - D	0.039	74.1	** +
9- innovation	forces	0.004	62.4	** +
	distorsions	0.141	80.0	
	normes + T	0.234	87.7	
	normes + D	0.462	86.6	
	normes - T	0.000	75.0	**
	normes - D	0.001	73.4	** +

ANNEXE D: Résultats sur la fiabilité des questions de préférence

(Comparaison entre les premier et deuxième tests.)

dimensions	quest.	$\bar{d}$	T	$T^{*}_{0.95}$	signif. à 95%
1- négociation	1	0.57	3.82	1.45	*
4- soutien	2	0.57	3.82	1.45	*
7- accomplissement	3	0.94	4.61	1.40	*
2- écoute	1	0.60	4.03	1.40	*
5- direction	2	0.86	4.66	1.34	*
8- perfection	3	0.91	4.43	1.52	*
3- partenariat	1	0.65	4.45	1.43	*
6- contrôle	2	0.51	3.90	1.38	*
9- innovation	3	1.03	4.40	1.47	*

Le défi posé par les nouvelles technologies.[1]

Sylvain Perron[2]

1.0 INTRODUCTION

L'arrivée sur le marché de nouvelles technologies de plus en plus perfectionnées amène beaucoup de changement dans les méthodes d'enquêtes. Alors que la collecte des données peut se faire de façon plus efficace, d'autres problèmes surgissent. À Statistique Canada, pour tirer pleinement profit de ces nouvelles technologies, il faut modifier ou adapter les méthodes de collecte dans les enquêtes tout en s'assurant du maintien de la confidentialité des données.

Par nouvelles technologies, on entend principalement: la technologie des téléphones à tonalité (touchtone), l'utilisation des micro-ordinateurs, les systèmes d'entrevues assistées par ordinateurs, la technologie de reconnaissance de la voix, la technologie de reconnaissance de l'écriture (gridpad) et les bélinographes ou télécopieurs. Ce sont les nouvelles possibilités qu'elles offrent en terme de coûts et d'amélioration de la qualité des données qui nous amènent à repenser nos méthodes d'enquêtes.

Le présent document énoncera certains problèmes ou certaines préoccupations soulevés par l'utilisation de ces nouvelles technologies. Il n'offre pas de solutions précises puisqu'il y a encore beaucoup de recherche à effectuer dans ce domaine. On soulève plutôt les problèmes, en esquissant des solutions potentielles.

Le document est divisé en quatre parties. En premier lieu, on verra les impacts et problèmes qui sont causés, au niveau des interviewers et des répondants, par l'utilisation des nouvelles technologies. Par la suite au chapitre trois, on décrira certains des défis méthodologiques qui doivent être relevés. Au chapitre quatre on discutera de trois études entreprises par Statistique Canada qui utiliseront des nouvelles technologies. Enfin, on terminera avec une brève conclusion.

---

1 Présenté au colloque sur les applications de la statistique du 58e congrès de l'Association canadienne française pour l'avancement des sciences, Québec, mai 1990.

2 Sylvain Perron, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Immeuble R-H Coats, Ottawa (Ontario), K1A 0T6.

2.0 PROBLÈMES CAUSÉS PAR L'UTILISATION DES NOUVELLES TECHNOLOGIES

Il est possible de considérer l'impact causé par l'utilisation des nouvelles technologies à deux niveaux. On peut s'intéresser aux répercussions qui affectent les interviewers ou à celles qui affectent les répondants. On doit aussi s'attarder à la nécessité de préserver la confidentialité de l'information tout au long des activités de collecte et de saisie, et en particulier lors de la transmission de ces données. On verra donc ces trois points plus en détails.

2.1 Impact au niveau de l'interviewer

Les différents problèmes qui affectent les interviewers ont une connotation très terre-à-terre. Ils sont souvent issus du matériel que les interviewers doivent utiliser, ou encore des conditions de travail dans lesquelles ils doivent opérer. On peut identifier deux types de problèmes.

Un premier type est relié au limite physique des machines que les interviewers utilisent lors d'entrevues. Le poids, la taille, l'éclairage requis pour pouvoir opérer ces machines sont déterminants. Une machine inadéquate posera surement des difficultés pour les interviewers lors de son utilisation. Dans des études menées en Suède [1] et aux États-Unis, on a montré que le poids des machines qui sont utilisées pour conduire des entrevues personnelles est un facteur majeur pour les interviewers. Si la machine est trop lourde, il y a un impact négatif sur les interviewers lors de son transport par exemple. Si elle est difficile d'utilisation (clavier trop petit, écran mal éclairé, période d' opération des piles trop courte,...) alors la machine devient une nuisance pour les interviewers.

Un second type de problèmes est relié à la façon dont on peut utiliser les nouvelles technologies. Un micro dont le clavier requiert que l'interviewer soit assis pour l'utiliser représente une embûche pour celui-ci lors d'entrevues personnelles où le répondant ne lui offre pas la possibilité d'entrer et de s'asseoir. Dans un environnement d'interview téléphonique assistée par ordinateur (ITAO), les problèmes de santé reliés aux yeux et au dos, peuvent être rencontrés plus fréquemment.

Un autre problème relié à l'utilisation du matériel s'applique surtout dans le cas des entrevues personnelles assistées par ordinateur. Les interviewers ont souvent le sentiment qu'en utilisant le micro, ils perdent le contact visuel et personnel qu'ils établissent généralement avec les répondants. Ce contact qui semble aisé à établir lorsque les interviewers utilisent un questionnaire papier, ne s'établit pas lorsqu'ils utilisent la machine.

Enfin, il faut aussi analyser les nouvelles tâches que doivent effectuer les interviewers pour pouvoir utiliser les nouvelles technologies. Si on requiert de ceux-ci qu'ils doivent recharger les piles, sauvegarder eux-même les données, relier la machine à un modem ou à une ligne téléphonique pour effectuer la transmission des données, on leur impose une plus grande responsabilité, un nouvel environnement de travail.

Ce nouvel environnement de travail apporte certains points négatifs. Par exemple, lorsque les interviewers sont appelés à travailler avec un système ITAO, ou encore lorsqu'ils sont responsables d'un micro-ordinateur, plusieurs études montrent que le niveau de stress qu'ils subissent est augmenté, ce qui cause un certain malaise [2]. On a aussi montré que les interviewers qui font des visites personnelles avec des micro-ordinateurs portatifs, ne se sentent pas en sécurité dans certaines situations (comparativement à l'utilisation des questionnaires sur papier).

Tous les points mentionnés jusqu'ici indique donc clairement qu'il faut tenir compte des interactions qui vont exister entre les interviewers et les nouvelles technologies. Malgré les aspects négatifs possibles, il y a beaucoup d'avantages à utiliser les nouvelles technologies. La conversion des méthodes de collecte vers l'utilisation des technologies de pointe peut avoir plusieurs impacts positifs pour les interviewers si l'introduction est faite correctement. Si leur travail est facilité par ces technologies, les interviewers les acceptent bien. Cette période d'acceptation sera d'autant plus courte que l'on fera sentir aux interviewers que les machines ne sont que des outils de travail destinés à faciliter leurs tâches.

Dans le cas d'entrevues personnelles assistées par ordinateur, il faut aussi s'assurer que les interviewers soient bien formés pour faire face aux problèmes qui peuvent survenir avec le micro puisque les interviewers n'ont pas accès aux spécialistes en informatique disponibles dans les bureaux. Une bonne formation va permettre aux interviewers d'acquérir les connaissances spécialisées en informatique, dactylographie ou autres, qui sont nécessaires afin de bien utiliser les nouvelles technologies. Un bon interviewer avec les technologies courantes, demeure un bon interviewer lorsque de nouvelles technologies sont utilisées. Donc, lorsque les nouvelles technologies sont bien amenées, on obtiendra la collaboration des interviewers.

2.2 Impact au niveau du répondant

Au niveau des répondants, on doit faire face à deux problèmes. En premier lieu, le fait d'utiliser les ordinateurs directement chez le répondant, ou encore grâce à un ITAO, entraine une prise de conscience de l'existence de fichiers électroniques concernant le répondant. Présentement ce fait ne cause pas de problèmes réels d'après les nombreuses expérience faites en Suède, en Hollande aux États-Unis ou encore au Canada. Cependant le syndrome du "Big Brother" doit demeurer présent dans l'esprit des méthodologistes.

L'utilisation des nouvelles technologies peut aussi entrainer une intrusion chez lui si par exemple l'interviewer doit s'asseoir pour utiliser sa machine, ou encore, s'il doit utiliser une prise de courant dans la maison. Il faut prévoir une utilisation qui n'affecte pas le répondant.

Par contre, un avantage non-négligeable qui est amené par ces technologies est la possibilité d'adapter les méthodes de collecte aux besoins et aux possibilités du répondant. Ainsi, l'utilisation de système ITAO avec son système de planification automatique des appels, permet de contacter les répondants au moment où ils le désirent, soit lorsque leurs données sont

disponibles, ou à un moment fixé par entente préalable. Ceci permet de réduire le fardeau des répondants puisqu'on ne les contacte que lorsqu'ils sont prêts à fournir leurs données.

D'autres possibilités comme l'utilisation des télécopieurs, du courrier électronique, ou encore la possibilité d'envoyer au répondant une disquette contenant le questionnaire ainsi que le programme nécessaire pour y répondre, ouvrent la voie à une nouvelle philosophie dans les méthodes de collecte des données. En général, ces nouvelles technologies offrent beaucoup d'avantages aux répondants, favorisant ainsi une meilleure collaboration.

Cependant, il ne faut pas oublier que les répondants ont aussi accès a ces nouvelles technologies. Ceci ajoute une autre pièce a un casse-tête déjà complexe. En effet, avec des services de gestion des appels (filtration des appels inopportuns par exemple) que la compagnie Bell Canada offre, il devient plus difficile de rejoindre les répondants. Les répondeurs téléphoniques et les téléphones cellulaires compliquent les enquêtes téléphoniques avec composition aléatoire de numéros de téléphone. Il faut donc adapter nos méthodes de collecte aux nouvelles technologies, mais aussi tenir compte de leurs utilisations par nos répondants.

2.3 Transmission des données

Le problème de la transmission des données est directement lié au mandat de Statistique Canada d'assurer et de protéger la confidentialité de l'information recueillie selon la loi de la statistique. Puisque c'est cette assurance de confidentialité qui facilite la collecte des données auprès des répondants, il ne faut pas la mettre en péril.

C'est pour cette raison par exemple qu'il existe à Statistique Canada deux réseaux informatiques. Le premier est un réseau public contenant entre autres le système de données CANSIM. Ce réseau ne contient que des données qu'on retrouvent dans les différentes publications de Statistique Canada.

Le second réseau par contre est confidentiel. Il contient toutes les données confidentielles recueillies lors des différentes enquêtes ou encore, lors du recensement. Pour en assurer la protection, il n'existe aucun lien externe permettant l'accès aux ordinateurs de ce réseau. Quand aux accès provenant de l'intérieur des locaux de Statistique Canada, ils sont protégés par un système de protection connu sous l'acronyme "ACF2". Ces procédures assurent la confidentialité et l'inviolabilité des données puisque les seules personnes pouvant accéder aux données sont des personnes autorisées par Statistique Canada, à partir des locaux de Statistique Canada situés à Ottawa.

Cet impératif de protéger l'information amène plusieurs restrictions sur la façon dont certaines nouvelles technologies peuvent être utilisées. Par exemple, lorsque des données doivent être transmises par lignes téléphoniques, elles doivent être codées afin d'être incompréhensibles si quelqu'un non-autorisé les intercepte. De plus, ce codage ne doit pas être fait à l'aide d'un programme (qu'on peut oublier d'exécuter), il doit être intégré à l'appareil qui effectue la transmission.

Ce même impératif a aussi limité par le passé l'utilisation des courriers électroniques ou des télécopieurs, par les répondants qui le désiraient. Il y a eu cependant un important changement dans la politique de Statistique Canada. Cette modification majeure, qui a pris effet au début de 1990, permets sur une base d'essai, l'utilisation de trois types de transmissions électroniques des données par les répondants. Ces trois types sont: un service commercial de courrier électronique, un service commercial de réseau à commutation par paquets, et enfin l'utilisation du télécopieur.

Pour pouvoir se prévaloir de cette politique, on demande aux répondants de signer une lettre d'entente relative à la collecte électronique. Par cette lettre, les répondants reconnaissent qu'ils désirent que leurs données soient recueillies par un de ces trois modes. Le répondant reconnait aussi les possibilités d'accès par des personnes non autorisées lors du transit des données. Enfin, la condition la plus importante est que le répondant relève Statistique Canada de l'obligation d'assurer la confidentialité des données, jusqu'à ce qu'un accusé de réception des données ne lui soit envoyé.

De cette façon, on s'assure que le répondant comprends bien tous les dangers potentiels qui existent avec cette forme de collecte. Cela permet quand même d'offrir des services qui accommodent les répondants. De plus, on indique comment Statistique Canada s'acquitte de son mandat.

Cette approche peut sembler trop prudente, et surtout très lente pour s'adapter aux technologies existantes. Cependant elle est guidée par le mandat de Statistique Canada, et par le besoin de maintenir la collaboration des répondants. Puisque cette collaboration est assurée en bonne partie par l'assurance de protection des données, il est impératif que la confidentialité et l'inviolabilité des données soient préservées.

3.0 DÉFIS À RELEVER

Comme on l'a mentionné, l'utilisation des nouvelles technologies permet entre autre de modifier les méthodes de collecte. Cependant, afin de prendre réellement avantage de celles-ci, plusieurs défis méthodologiques doivent être relever. Dans cette section, on verra quatre aspects méthodologiques où de la recherche et du développement sont requis.

3.1 Développement de systèmes génériques

On doit mentionner l'effort majeur que Statistique Canada a investi dans le développement de fonctions d'enquête génériques. En effet, toute enquête comporte en général les mêmes étapes soit la sélection de l'échantillon, la collecte et la saisie, la vérification et l'imputation, et l'estimation. Étant donné la technologie informatique qui existait auparavant, on devait développer des systèmes spécifiques pour chacune des enquêtes. Ce développement de multiples systèmes représentait un gaspillage de ressources humaine et matérielle puisqu'on dupliquait souvent les mêmes fonctions [3].

L'avènement des micro-ordinateurs très performant, et le développement des réseaux, rend possible le développement de systèmes génériques qui, avec un minimum de modification, peuvent être utilisés par chacune des enquêtes. Ceci représente une économie de ressources tout en permettant de mettre l'accent sur la recherche de solutions à certains problèmes méthodologiques qui ne peuvent être résolus présentement, à cause d'un manque de ressources.

Actuellement, il existe quatre systèmes génériques en cours de développement. Ce sont : la fonction générique d'échantillonnage, la fonction générique de collecte et de saisie des données, la fonction générique de vérification et d'imputation, et enfin, la fonction générique d'estimation. Lorsque ces fonctions seront toutes opérationnelles (la fonction de vérification et d'imputation est déjà opérationnelle), il sera alors possible d'effectuer une enquête complète par leurs entremises.

3.2 Environnement de travail interactif

On a assisté à la transformation de l'environnement de travail. On est passé très rapidement d'un environnement de travail par lots des années 1970, à un environnement de travail interactif. Par exemple, on peut maintenant développer des méthodes de vérification interactive des données, au lieu de s'en tenir au système par lot. Le travail se fait plus rapidement et plus efficacement.

Il faut donc ré-évaluer les différentes méthodes utilisées par les enquêtes afin d'identifier les étapes qui bénéficieraient de ce nouvel environnement. Il se peut que certaines méthodes ont été rejetées auparavant, simplement parce que la technologie ne le permettait pas. Avec les changements qui sont survenus dans le milieu informatique, il faut les étudier de nouveau.

3.3 Utilisation de modes de collecte mixtes

Parmi les fonctions génériques d'enquêtes qui sont en développement, on peut mentionner la fonction générique de collecte et de saisie de l'information. Cette fonction offrira à un gestionnaire d'enquête le choix de plusieurs modes de collecte: par la poste, par téléphone, par interview téléphonique assistée par ordinateur, par entrevue personnelle, par télécopieur, etc... Ce qui est significatif des nouvelles technologies, c'est que le gestionnaire ne sera plus limité dans son choix des modes de collecte. On pourra vraiment parler de modes mixtes. Il ne s'agira plus d'une enquête par la poste avec suivi effectué par téléphone par exemple. Il s'agira d'une enquête où une partie de l'échantillon sera enquêtée par la poste, une autre par entrevue personnelle, une autre par télécopieur, etc.

Évidemment, la répartition de l'échantillon parmi les différents modes de collecte devra se faire de façon optimale. Divers facteurs devront être analysés. Le coût des différents modes est un facteur important. La durée de la période de collecte est aussi un facteur, de même que les erreurs et biais encourus par chacune d'elles. On doit aussi mentionner les besoins et capacités des répondants, si un répondant refuse de donner ses données par téléphone, il faudra alors l'assigner à un autre mode de collecte.

L'utilisation des modes mixtes devient possible surtout grâce au développement d'un bon système de gestion de l'information. Il faut en effet pouvoir identifier en tout temps, quelles unités de l'échantillon ont été traitées, quelles unités sont encore à traiter, etc. Ce système doit aussi gérer le transfert d'unités d'un mode de collecte à un autre, lorsque certaines dates limites sont atteintes. Par exemple, des unités par la poste seront transférées au téléphone si leurs questionnaires n'ont pas été reçus avant la dernière semaine de collecte.

En ce qui concerne les systèmes d'ITAO, un défi qui doit être relevé est le développement d'un bon planificateur d'entrevues. Ce planificateur devra ordonner tous les cas en attente selon leurs priorités, en tenant compte de multiples facteurs. Par exemple dans une enquête répétitive, il faut tenir compte des naissances, des unités ayant un rendez-vous fixes, des cas de signal occupé, etc. Il faut aussi pouvoir changer les priorités au fur et à mesure que l'enquête progresse, certaines unités peuvent devenir hautement prioritaires. Il faut tenir compte des unités qui sont transférée d'un autre mode de collecte. Il s'agit donc là d'un problème de files d'attentes dynamiques.

De plus, ce planificateur doit être conçu de façon à répartir les entrevues entre les interviewers pour minimiser et surtout, pour pouvoir évaluer le biais dû aux interviewers. Ce biais est défini par Groves et Magilavy [4] comme la tendance que les données des répondants ont de varier selon l'interviewer qui a a fait les entrevues. Par exemple, ce problème peut surgir dans une enquête nationale qui sera utilisée pour faire des estimations pour des petites régions. Il faut en effet s'assurer qu'aucun interviewer en particulier, ne s'acquitte de tous les cas qui appartiennent à une sous-région donnée; sinon le biais de l'interviewer pourra affecter les estimations de toute une sous-région. En tenant compte de cet aspect lorsqu'il alloue les cas, le planificateur permet de contrôler le biais, et de l'évaluer.

On voit donc que les nouvelles technologies offrent plusieurs opportunités. Certaines limites qui étaient imposées non pas du point de vue méthodologique, mais par les capacités technologiques, sont maintenant levées.

3.4 Conception de questionnaire

L'arrivée des modes de collecte mixtes et des systèmes d'interviews assistées par ordinateur amène des changements au niveau de la conception des questionnaires. Par exemple, les modes de collecte assistés par ordinateur permettent la création d'un questionnaire plus personnalisé. L'ordinateur va se charger d'inclure dans les questions le nom du répondant, ou va utiliser le bon genre (féminin ou masculin) pour les questions.

Si on analyse plus en détail les systèmes assistés par ordinateur, on s'aperçoit qu'un des avantages qu'ils nous apportent est la possibilité d'avoir des questionnaires à branchements complexes. En effet, comme les branchements sont déterminés par la machine, et non pas par l'interviewer, il y a moins de risque d'erreurs de cheminement dans le questionnaire.

L'ordinateur doit aussi conserver en mémoire le cheminement qui a été suivi dans le questionnaire. Ceci va permettre à l'interviewer de revenir en arrière afin de vérifier des réponses douteuses ayant échouées des règles de vérification. Il faut que ces mouvements non-conventionnels à travers le questionnaire se fassent facilement. Le système doit gérer le mouvement vers la question fautive, puis ramener l'interviewer au point où l'entrevue était rendue.

La théorie des graphes semblent nous permettre de surmonter ce problème de branchements et de mouvements non-conventionnels [5]. Elle nous offre les outils qui seront nécessaires pour programmer correctement le système, afin de tenir compte de ces contraintes. Malgré tout, il faut développer ces outils ou ces programmes.

La théorie des graphes va aussi permettre de contrer un autre problème que les interviewers rencontrent lorsqu'ils utilisent un système assisté par ordinateur. Ce problème est celui de la segmentation [6], qui surgit parce que l'interviewer n'a plus une vue d'ensemble du questionnaire. Contrairement à un questionnaire sur papier, où l'interviewer voit dans quelle section, ou à quel niveau il est rendu dans l'entrevue, un questionnaire électronique ne lui fournit pas cette information. Pour contrer le problème, la théorie des graphes pourrait permettre la création automatique d'un schéma global du questionnaire, indiquant sommairement à l'interviewer où il en est dans l'entrevue.

En ce qui concerne les modes de collecte mixtes, le problème qui doit être résolu est de savoir si un seul questionnaire peut être utilisable par plusieurs modes, ou s'il faut des questionnaires spécifiques à chaque mode. La cohérence de l'information, le biais qui peut être engendré par les types de questionnaires, les variances qui peuvent être différentes, tous ces points doivent être analysés plus en détails. Il n'est pas évident qu'un seul questionnaire soit suffisant pour tous les modes. Et si on ne peut pas changer de questionnaires pour utiliser pleinement la capacité de chacun des modes, ceci aura des répercussions sur l'utilisation potentielle de ces modes.

4.0 TROIS ÉTUDES À STATISTIQUE CANADA

Afin d'évaluer des nouvelles technologies pour la collecte des données, Statistique Canada a adopté une approche en trois phases. Une étude peut évaluer la technologie elle-même, c'est-à-dire qu'on va analyser si la technologie existante peut répondre aux besoins d'une enquête. Dans une seconde phase, on peut vouloir évaluer non seulement la technologie, mais aussi les méthodes. À ce moment, on veut regarder de quelle façon les méthodes de collecte peuvent être modifiées pour profiter de la technologie. Ce peut être par exemple par l'addition de vérifications interactives.

Enfin, la troisième phase est un test complet et intégré, où on fera un test de masse de toutes les opérations de collecte qui seraient modifiées par la nouvelle technologie. Ce n'est qu'après cette troisième phase qu'une enquête serait convertie vers un nouveau mode de collecte.

Au cours de l'année 1990, Statistique Canada a entrepris trois études concernant l'utilisation de trois nouvelles technologies soit: le gridpad, le touchtone, et un micro-ordinateur portatif-manuel. Les études du gridpad et du touchtone correspondent à l'évaluation de la technologie seulement; l'étude de micro-ordinateur portatif manuel correspond à l'évaluation des méthodes et de la technologie.

Grâce à ces trois études, Statistique Canada va acquérir l'expertise méthodologique, informatique et opérationnelle requise pour l'utilisation de ces technologies. Il sera aussi possible d'évaluer les économies potentielles pouvant être réalisées en utilisant ces technologies, en particulier si on supprime la saisie de questionnaires. Ces études permettront aussi d'évaluer l'impact de ces trois technologies sur les interviewers.

De plus, l'utilisation des nouvelles technologies devraient permettre une diminution du temps requis pour le traitement des enquêtes. Cette utilisation devrait améliorer les conditions de travail, et par conséquent diminuer le stress du personnel. Il y aura aussi plus de temps disponible pour la révision et l'évaluation des enquêtes, au lieu d'effectuer des tâches plus routinières de collecte et de saisie. Enfin, on peut mentionner que les résultats de l'enquête pourraient être disponibles plus rapidement.

4.1 Utilisation d'un micro-ordinateur "GRIDPAD"

Cette première étude sera effectuée pour l'enquête trimestrielle sur le camionnage pour compte d'autrui (E.T.C.C.A.). Le travail des interviewers dans cette enquête consiste à se rendre chez des entreprises de camionnage, à sélectionner des factures parmi toutes celles de l'entreprise, et de saisir l'information qui se trouve sur les factures soit, le coût du transport, l'origine et la destination de la marchandise transportée, ainsi que la marchandise transportée.

Présentement, cette opération se fait manuellement à l'aide d'un formulaire de papier. Le but de l'étude est de tester si un ordinateur de type Gridpad peut être utilisé pour effectuer la saisie des données. Un gridpad est un ordinateur qui possède un écran sur lequel on écrit à l'aide d'un crayon spécial. L'ordinateur est capable de reconnaître l'écriture. On peut aussi utiliser des questions à choix multiples, où les interviewers n'ont qu'à cocher la bonne réponse.

Le test servira surtout à vérifier que le temps d'apprentissage des interviewers n'est pas trop long. En effet, ceux-ci doivent ré-apprendre à écrire de sorte que l'ordinateur puisse reconnaitre leur écriture. De plus, on veut vérifier qu'il n'y a pas de problèmes à utiliser le gridpad dans les divers endroits de travail. Les conditions de travail sont très diverses et peuvent causer certains problèmes.

4.2 Utilisation de la technologie "TOUCHTONE"

C'est avec les petites entreprises de l'enquête sur emploi la rémunération et les heures de travail (E.E.R.H.) que le touchtone va être évaluer. Il s'agira ici aussi d'une évaluation de la technologie qui existe présentement dans ce domaine. Le système touchtone consiste en un ordinateur qui reconnait les impulsions émises par un téléphone touchtone.

En général, lorsqu'une enquête utilise ce mode de collecte en conjonction avec d'autres, les répondants reçoivent un horaire leur indiquant la période de collecte durant laquelle on désire qu'ils rapportent leurs données. À partir de là, le répondant fournit ses données quand il le désire. Il contacte le système, s'identifie à l'aide de son numéro d'identification. Il fournit alors ses données dans l'ordre requis par le système, en utilisant les touches de son téléphone à tonalité.

Au cours d'étude similaires aux États-Unis, on a montré que cette technologie pouvait être introduite facilement auprès de répondants qui font déjà parti d'un échantillon répondant par ITAO. L'étude qui est entreprise vérifiera si cette conclusion tient aussi au Canada parmi les répondants qui sont contactés par téléphone. On vérifiera aussi l'attitude des répondants par la poste.

Au niveau de l'utilisation du touchtone, il reste encore plusieurs questions en suspens. Par exemple, il faut vérifier le maintien du taux de réponse à l'intérieur de l'échantillon du touchtone. Il semble en effet qu'il ne suffit pas d'envoyer un calendrier global des périodes de collecte pour assurer des réponses à jour. Il faut aussi envoyer des cartes de rappel à chaque période de collecte. Cette procédure diminue les gains potentiels qui peuvent être réalisés avec le touchtone.

Il faudra aussi vérifier la possibilité d'effectuer des vérifications interactives avec le touchtone. En plus de regarder du côté du potentiel de la technologie, il faut aussi se pencher sur l'impact de ces vérifications sur le répondant.

4.3 Utilisation d'un micro-ordinateur portatif manuel

Le troisième type de technologie qui est testé par Statistique Canada, est le micro-ordinateur portatif manuel. On se servira de l'enquête de l'indice des prix à la consommation (IPC) pour évaluer cette technologie. Cette évaluation portera sur la technologie elle-même, mais aussi sur les méthodes. L'évaluation est donc à un niveau plus avancé que pour les deux premières études.

La collecte des données pour l'IPC consiste à ramasser les prix d'un certain nombre d'items qui composent un panier de biens et services typiques consommés par une famille canadienne moyenne. Parmi ces items on retrouve des biens d'épicerie, des vêtements, ou encore des services comme des repas pris au restaurant ou les taxes municipales. À chaque mois les interviewers vont saisir les prix d'articles et de services pré-déterminés. Ainsi, les articles

d'épicerie sont effectués deux fois par mois, et représentent 80% de la charge de travail mensuelle. Par contre, les articles de sports d'été ou d'hiver sont effectués durant la saison appropriée.

Pour effectuer la collecte, les interviewers reçoivent chaque mois des formulaires pré-imprimés sur lesquels se trouve les items et services dont ils doivent saisir les prix. Il y a deux types de formulaires. Le premier est un formulaire standard généré par ordinateur. Ce formulaire existe en une seule version, et il est utilisé pour saisir l'information de 95% des biens et services du panier de consommation. Le second formulaire est de type non-standard. Il en existe plusieurs versions pour saisir différents items. Par exemple, il y a le formulaire pour les voyages en taxi, celui pour les voyages en autobus, celui pour les consommations alcooliques. Toutes ces versions sont différentes et elles ne sont pas générées automatiquement par ordinateur.

Parce que le second type de formulaire est complexe, qu'il possède plusieurs versions, et qu'il ne représente qu'un faible pourcentage de la charge de travail des interviewers, il ne sera pas remplacé par le micro. C'est le formulaire standard qui sera remplacé par la machine. L'étude se concentre donc sur les opérations qui sont reliées à ce formulaire. De plus, l'étude sera restreinte à la composante épicerie de l'IPC, puisqu'elle est représentative du travail effectué à l'aide du formulaire standard.

Trois objectifs globaux ont été établis pour l'étude. En premier lieu, on veut évaluer la faisabilité d'effectuer la collecte de la composante épicerie de l'IPC au moyen du micro. On veut aussi effectuer une analyse de la qualité des données recueillies par le micro. Enfin, on analysera la possibilité d'effectuer des vérifications interactives au moyen du micro.

Présentement, ces vérifications interactives sont effectuées au niveau de la saisie dans les bureaux régionaux. Lorsque des erreurs sont identifiées, il est difficile de les corriger rapidement. En effectuant les vérifications lorsque l'interviewer est encore dans le point de vente, il sera facile de corriger ces erreurs.

Concernant le premier objectif, en plus de vérifier si techniquement le micro-ordinateur manuel peut être utilisé pour effectuer la collecte, on veut aussi s'assurer de sa maniabilité et de sa fonctionnalité auprès des interviewers. On désire aussi identifier les aspects de la collecte qui sont appelés à être modifiés, afin de mieux utiliser la puissance du micro. On vérifiera en plus la performance du micro en terme de sa capacité de rétention de l'information recueillie par les interviewers.

Un aspect important de l'utilisation du micro-ordinateur portatif manuel qui ne sera pas évalué au cours de cette étude est la transmission des données entre les interviewers et les bureaux régionaux. Ces transmissions comprennent l'envoi à l'interviewer de l'information nécessaire pour effectuer la collecte des prix, et le retour au bureau régional de l'information recueillie. Cet aspect sera évalué lors d'une prochaine étude, si les résultats du premier test sont concluants.

5.0 CONCLUSION

En conclusion, on peut mentionner que pour tirer pleinement profit des nouvelles technologies, il faut faire plus que de remplacer un questionnaire papier par l'ordinateur. Il faut aussi modifier les méthodes de collecte. Ceci ne peut se faire qu'en collaboration entre méthodologistes, informaticiens et spécialistes des sujets. Il faut aussi analyser les impacts que ces technologies ont au niveau des répondants et des interviewers afin de bien les introduire. Il reste encore beaucoup de recherche à faire afin de bien appliquer ces technologies [7].

Au niveau de Statistique Canada, l'adoption de nouvelles technologies se fait lentement. Comme il a été mentionné auparavant, cette approche prudente est liée au mandat de Statistique Canada d'assurer la confidentialité et l'inviolabilité des données. Ces deux contraintes obligent une analyse très poussée de toutes les implications qu'apportent une nouvelle technologie avant son utilisation.

RÉFÉRENCES

[1] CADAC Project, "Computer-assisted data collection in the labour force survey, report on technical tests", Statistics Sweden, janvier 1989.

[2] Catlin G., Ingram S., Hunter L., "The effect of CATI on data quality and costs: a comparison of CATI and paper methods", CATI research project, document interne, Statistique Canada, 1988.

[3] ACFAS, Recueil des textes des présentations du colloque sur les méthodes et domaines d'application de la statistique, Bureau de la statistique du Québec, 1988.

[4] Groves R. M., Magilavy L. J., "Measuring and explaining interviewer effects in centralized telephone surveys", Public Opinion Quaterly, vol. 50, 1986.

[5] Willenborg, L.C.R.J., "The routing structure of questionnaire", document interne, Central Bureau of Statistics, Netherlands.

[6] House C., "Questionnaire design with computer assisted telephone interviewing", Journal of Official Statistics, vol.1, no.2, Statistics Sweden, 1985.

[7] Nicholls II, W.L. and Groves R.M., "The status of computer assisted telephone interviewing Part I and Part II", Journal of Official Statistics, vol.2, no.2, Statistics Sweden, 1986.

Analyse statistique des effets de la nature des aliments sur l'activité myoélectrique réticulo-ruminale et sur le comportement chez le veau

Anick Bérard*
Station de recherches, Agriculture Canada,
Lennoxville, Québec, J1M 1Z3

RÉSUMÉ

Trois types d'alimentation ont été donnés à six veaux pendant trois périodes de trois semaines pour observer l'effet de diète sur l'activité myoélectrique réticulo-ruminale et le comportement des veaux. Les animaux ont été distribués aléatoirement selon un modèle "crossover" et la prise de données a été faite au cours de deux phases (24 heures, 48 heures) à l'aide de cassettes vidéo et d'un physiographe. Le comportement des données de digestion de chaque phase a été estimé par des courbes de Fourier pour pouvoir déterminer la durée des cycles de l'activité myoélectrique. La méthode d'analyse de Rowell et Walters (1976) a été employée en utilisant comme contrastes des courbes de Fourier pour définir le degré de la courbe qui s'ajustait le mieux à chaque variables mesurées.

INTRODUCTION

De plus en plus d'expériences en agriculture étudient l'activité du système digestif et le comportement chez les animaux. L'activité myoélectrique est un des indices du nombre de contractions et des mouvements ayant lieu dans l'appareil digestif. Des études antérieures ont déjà démontré que les mouvements dans le rumen pendant la digestion étaient reliés à la fréquence des contractions et que l'activité réticulo-ruminale était influencée par les caractéristiques physiques des aliments comme le type et la quantité de fibres. De plus, il a été prouvé que le type de diète modifiait l'activité myoélectrique et qu'il existait une relation directe entre la diminution d'activité réticulo-ruminale et le sommeil paradoxal.

L'expérience a été menée principalement pour observer l'effet de diète et l'effet du changement de diète sur l'activité réticulo-ruminale et sur le comportement

* L'expérience a été menée par le Dr. C. Girard et le Dr. S. Robert.

des veaux. En premier lieu, la méthodologie de l'expérience sera exposée. Par la suite, les méthodes statistiques et quelques résultats seront démontrés.

MÉTHODOLOGIE

Trois alimentations (foin & moulée séparés, ensilage & moulée en ration totale mélangée et ensilage & moulée séparés) ont été données à six veaux. Ces trois types de diète sont régulièrement utilisées sur les fermes. L'expérience a duré 9 semaines pendant lesquelles chaque veau a reçu un nouveau traitement à toutes les trois semaines (à toute les périodes). Les données ont été recueillies en deux phases : la phase adaptation (le premier 24 heures de la période) et la phase diète (2 X 24 heures à partir du $14^{ième}$ jour de la période).

Pour la collecte des données de digestion, trois groupes d'électrodes ont été implantés par chirurgie sur la paroi externe du rumen ainsi que trois autres sur la paroi externe du réseau. Au moment de la collecte, chaque électrode était reliée à un physiographe qui permettait de recueillir le nombre de contractions dans le rumen et dans le réseau. Les données ont été prises en continu mais pour l'analyse, le total par heure a été utilisé. En ce qui a trait aux données de comportement, elles ont été enregistrées sur cassettes vidéo. Pendant chaque phase, chaque veau a été observé et son comportement noté. Le comportement a été par la suite codé et la fréquence par heure utilisée à des fins d'analyse.

MÉTHODES STATISTIQUES

Six veaux agés de 23 semaines ont été distribués aléatoirement selon un modèle "crossover" pour étudier les effets de trois diètes. A chaque phase de chaque période, les veaux ont été utilisés comme unités expérimentales. Les facteurs du modèle ont été les périodes (1, 2 et 3), les veaux (1,..,6) et les traitements (A, B et C)(tableau 1). Les variables mesurées sont énumérées dans le tableau 2. Le nombre de contractions dans le réseau et dans le rumen a été décomposé de façon à faire ressortir le trajet exact des aliments dans le système digestif. La liste des codes de comportement n'est pas décrite dans ce texte mais chaque comportement a été divisé en fréquences pour les 24 heures, pour

Tableau 1. Description du plan expérimental - crossover

	veaux					
périodes	1	2	3	4	5	6
1	A	B	C	C	A	B
2	C	A	B	A	B	C
3	B	C	A	B	C	A

2 phases/ traitement : - phase adaptation (24 hres)
- phase diète (2 x 24 hres)

Traitements : A- foin & moulée séparés
B- ensilage & moulée en R.T.M.
C- ensilage & moulée séparés

Facteurs : veaux
périodes
traitements

le jour, pour la nuit ainsi que pour les 2 heures avant et les 2 heures après les repas. Du fait qu'il existait une corrélation évidente entre les deux phases, cet effet a été mesuré par la différence relative entre la phase adaptation et la phase diète. Dans les cas où une différence significative a été détectée entre les traitements, on a pu conclure qu'il existait une différence entre les deux phases pour la variable étudiée.

Chaque variable a été analysée phase par phase et l'effet phase a été mesuré à l'aide de la différence relative selon le modèle

$$Y_{ijkl} = u + V_i + P_j + T_k + e_{ijkl} \qquad (1)$$

où i = 1,..,6 (veaux)
j = 1,2,3 (périodes)
k = A,B,C (traitements)
l = 1,..,24 (heures)

Une moyenne des deux jours a été faite pour l'analyse de la phase diète qui s'étendait sur deux cycles de 24 heures. Même si théoriquement la variabilité a été diminuée en utilisant la moyenne, ceci n'a changé en rien les résultats du fait que les données ont été enregistrées sur plusieurs heures et que seulement deux données ont été impliquées à chaque calcul.

Tableau 2. Variables mesurées

- # contractions dans le réseau
  - total
  - monophasique
  - biphasique
  - triphasique
- # contractions dans le rumen
  - total
  - type A
  - type B
- Effet phase - (P2-P1)/P1
- # comportements (plusieurs variables)
  - 24 hres
  - jour
  - nuit
  - 2 hres avant le repas
  - 2 hres après le repas

Un aperçu général du comportement des données dans le temps était également intéressant. Après avoir regardé l'allure des variables mesurées à chaque heure pour chaque phase, il a été jugé approprié d'estimer les courbes par des courbes de Fourier. Ceci nous a permis de déterminer les heures où l'activité digestive et comportementale des veaux a été maximale et les heures où elle a été minimale. La durée d'un cycle, c'est-à-dire le temps entre deux maximums ou deux minimums, a été perçue à l'aide des courbes de Fourier. Les courbes avaient la forme suivante,

$$Y = a_0 + a_1\cos(Ch) + b_1\sin(Ch) + a_2\cos(2Ch) + b_2\sin(2Ch) + \ldots + a_{11}\cos(11Ch) + b_{11}\sin(11Ch) + a_{12}\cos(12Ch) \qquad (2)$$

où h = heure, h = 0,..,23
C = $360^\circ/24 = 15^\circ$
a_0, a_1, b_1, ... , a_{12} = constantes à estimer

où les sources de variation Heure, Veau X Heure, Période X Heure, Traitement X Heure et Erreur X Heure étaient subdivisées en 12 éléments associés aux paires de termes d'une courbe de Fourier.

Pour avoir un bon ajustement des données originales, il a été important de déterminer quel degré de la courbe donnait la meilleure estimation. Déterminer si les données s'expliquaient avec une ou plusieurs courbes présentait également un grand intérêt. Pour ce faire, la méthode de Rowell et Walters (1976) a été employée. Cette méthode est appropriée pour analyser des données qui ont été prises à répétition sur chaque unité expérimentale provenant d'une même expérience. Cette méthode peut s'appliquer pour des données enregistrées dans le temps et dans les cas où une unité expérimentale est échantillonnée plusieurs fois au même moment. La méthode de Rowell et Walters (1976) consiste à

analyser des contrastes et à tester les effets du modèle avec leur erreur respective. Elle est toujours valide, les calculs sont simples et l'interprétation est directe. Dans notre cas, les contrastes utilisés ont été des courbes de Fourier (éq. 2). Il existait une courbe par heure pour chaque unité expérimentale. Dû au très grand nombre de contrastes par unité expérimentale (24 courbes), il a été décidé qu'une courbe de Fourier par unité expérimentale, donc par 24 heures, serait suffisante. La somme, terme par terme, de chaque courbe a été faite pour donner un contraste par unité expérimentale par 24 heures. Chaque terme de toutes les variables mesurées a été analysé selon le modèle "crossover" (éq. 1) et les deux sommes de carrés de chaque degré ont été additionnés pour faire ressortir la variance totale de la courbe de Fourier (tableau 3). L'analyse avec la méthode de Rowell et Walters (1976) ayant comme contrastes les degrés d'une courbe de Fourier est représentée dans le tableau 4. Tous les carrés moyens sont identiques à ceux des analyses faites degré par degré, excepté pour un diviseur calculé comme la somme des carrés des coefficients dans le contraste.

Tableau 3. Analyse des contrastes terme par terme.

Ui		Vi	
	Moyenne		Moyenne
	Veaux		Veaux
	Périodes		Périodes
	Traitements		Traitements

i=ième degré de la courbe de Fourier

$$UVi = \sum (SS_{ui} + SS_{vi})$$

pour chaque effet correspondant

Ui = cos(iCh)
Vi = sin(iCh)

RÉSULTATS

Pour illustrer les méthodes statistiques utilisées, une brève description des résultats de la variable "rumen total", indiquant le nombre moyen de contractions ruminales par heure, sera donnée. Les résultats de l'analyse "crossover" ont démontrés qu'il existait une différence significative au seuil de

Tableau 4. Analyse de la variance selon Rowell et Walters (1976) avec la décomposition des effets pour une courbe de Fourier.

VEAU
PERIODE
TRAIT
TEMPS
DEGRE 1
DEGRE 2
.
DEGRE 11
COS12
TEMPS*VEAU (X CONTRASTE)
DEGRE 1 X VEAU
DEGRE 2 X VEAU
.
DEGRE 11
COS12 X VEAU
TEMPS*PERIODE (X CONTRASTE)
TEMPS*TRAIT (X CONTRASTE)
TEMPS*ERREUR (X CONTRASTE)

10% entre les traitements A et C (foin & moulée séparés et ensilage & moulée séparés) dans les deux phases. Une alimentation à base de foin génère moins d'activité dans le rumen qu'une alimentation à base d'ensilage (P<.10). Aucun effet de phase n'a cependant été détecté au seuil significatif de 10% ; ce qui prouve que l'activité dans le rumen ne change pas au cours des deux phases. L'analyse de Rowell et Walters (1976) a indiqué qu'un modèle du $4^{ième}$ degré (i.e. 9 paramètres) serait suffisant (P<.01) pour décrire le phénomène étudié pour chacune des phases. Les courbes du $4^{ième}$ degré qui s'ajustent le mieux aux observations (figures 1 et 2) ont pour équations

phase 1: $R = 126,65 + 8,10\cos(Ch) + 7,63\sin(Ch) + 3,09\cos(2Ch) + 6,71\sin(2Ch) + 4,79\cos(3Ch) + 6,25\sin(3Ch) + 4,03\cos(4Ch) + 1,36\sin(4Ch)$ $\quad R^2 = 78,92\%$

phase 2: $R = 126,14 + 7,49\cos(Ch) + 8,21\sin(Ch) + 2,66\cos(2Ch) + 6,61\sin(2Ch) + 5,83\cos(3Ch) + 5,79\sin(3Ch) + 3,78\cos(4Ch) + 1,15\sin(4Ch)$ $\quad R^2 = 83,54\%$

où $h = 0,..,23$

$C = 360^o/24 = 15^o$

Les composantes des deux courbes ainsi que leur maximum atteint sont donnés dans le tableau 5. Le maximum d'une courbe de Fourier est calculé selon les équations exposées dans le tableau 6. Pour la phase deux, la première composante (cycle de 24 heures) atteint son maximum à 3h11 (figure 4, cycle 1), la deuxième atteint son maximum à 4h32 (figure 4, cycle 2), la troisième composante atteint son maximum à 2h59 (figure 4, cycle 3) et la quatrième composante atteint son maximum à 1h08 et à toutes les 6 heures suivantes (figure 4, cycle 4). L'analyse de Rowell et Walters (1976) pour la phase un détectait ($P<.01$) une interaction entre les temps et les traitements. Il a donc été intéressant d'ajuster une courbe de Fourier pour chaque traitement. Dans la figure 5, on peut voir qu'effectivement une différence existe entre les courbes.

Tableau 5. Maximum atteint pour chaque composante des courbes de Fourier.

	Phases	
	1	2
Composantes	maximums	maximums
1	2h53	3h11
2	4h21	4h32
3	3h30	2h59
4	1h14	1h08

1 (24 hres), 2 (12 hres), 3 (8 hres), 4 (6 hres)

CONCLUSION

Lorsque l'on planifie une expérience selon un modèle "crossover", il est important de s'assurer qu'aucun effet résiduel ne s'accumule entre les

traitements. Si cela ne peut être évité, d'autres modèles sont plus appropriés pour ce genre de données, tels les carrés latins. La méthode de Rowell et

Tableau 6. Équations pour le calcul des maximums.

δ = angle où la "vague" atteint son maximum

$\delta' = \arctan (b_i / a_i)$

$b_i > 0$ et $a_i > 0 \rightarrow \delta = \delta'$

$b_i > 0$ et $a_i < 0 \rightarrow \delta = 180° - \delta'$

$b_i < 0$ et $a_i < 0 \rightarrow \delta = 180° + \delta'$

$b_i < 0$ et $a_i > 0 \rightarrow \delta = 360° - \delta'$

Walters (1976) est très efficace pour analyser des données prises dans le temps. Chaque effet est testé avec son erreur correspondante par opposition à d'autres analyses qui sont fréquemment utilisées (split-plot) et qui ne tiennent pas compte de la corrélation qui existe entre les données.

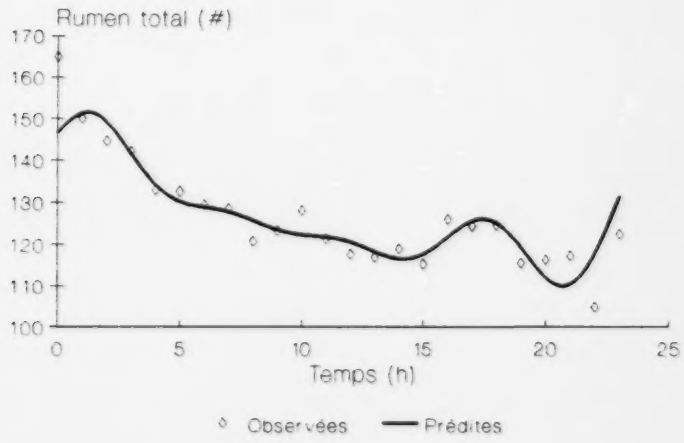


Figure 1. Courbe de Fourier de degré quatre - phase 1

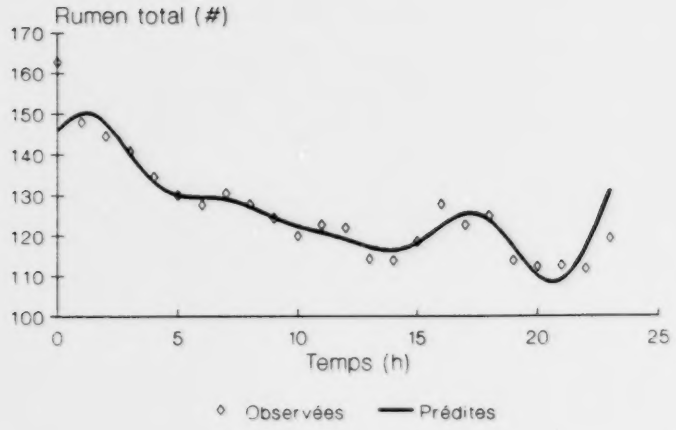


Figure 2. Courbe de Fourier de degré quatre - phase 2

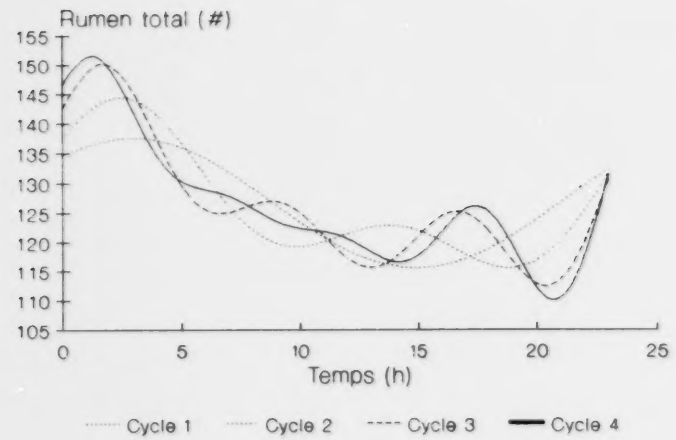


Figure 3. Composantes - courbe de Fourier - phase 1

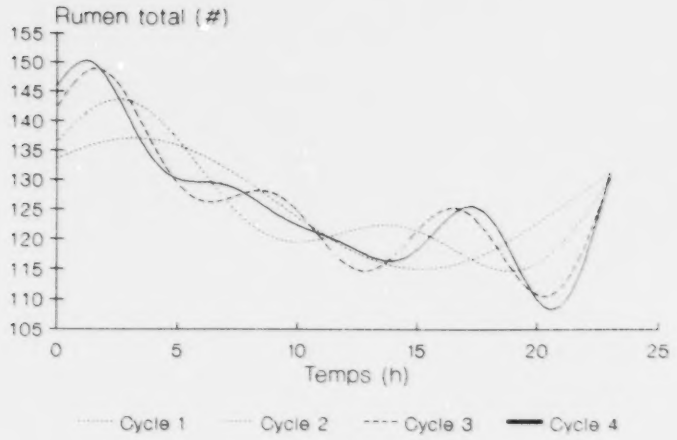


Figure 4. Composantes - courbe de Fourier - phase 2

RÉFÉRENCES

HILLS, M. et P. ARMITAGE.(1979). The two period crossover clinical trial. *Br.J.Clin.Pharmac.* **8**, 7-20.

LITTLE, T.M. et F.J. HILLS.(1978). *Agricultural Experimentation, Design and Analysis*. John Wiley & Sons, Toronto.

MONTGOMERY, D.C.(1984). *Design and Analysis of Experiments*, 2nd ed. John Wiley & Sons, Toronto.

NETER, J., W. WASSERMAN et M.H. KUTNER.(1985). *Applied Linear Statistical Models*, 2nd ed. Irwin, Homewood, Illinois.

ROWELL, J.G. et D.E. WALTERS.(1976). Analysing data with repeated observations on each experimental unit. *J. Agric. Sci.* **87**, 423-432.

SAS.(1986). *SAS User's Guide: Statistics*. SAS Inst., Inc., Gary, NC.

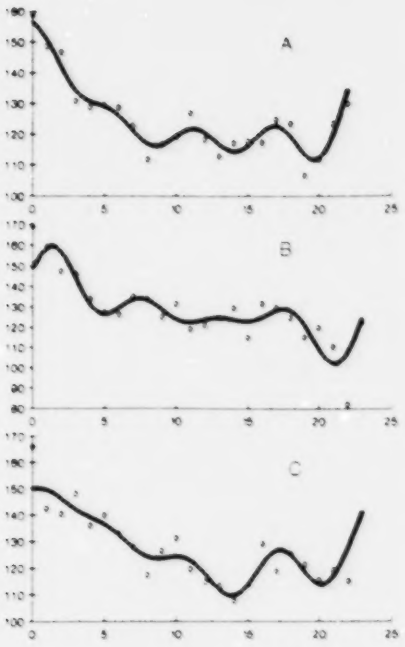


Figure 5. Courbes de Fourier par traitement - phase 1
(# contractions dans le rumen - total)

Analyse statistique des préférences des Canadiens en matière de bière

Jean-Pierre Carmichael, Gaétan Daigle et Louis-Paul Rivest
Université Laval

1. INTRODUCTION

Les grandes brasseries canadiennes organisent régulièrement des sessions d'évaluation de leurs bières, des bières de leurs concurrents ou encore des bières qu'elles sont en train de développer. Des individus de différentes populations-cibles participent à ces sessions.

Au cours d'une session, un individu goûte plusieurs bières présentées en petite quantité et dans des conditions identiques. Il donne à chaque bière une cote d'appréciation sur une échelle de 1 (déteste beaucoup) à 9 (aime beaucoup).

Dans le cadre des études de cas de la Revue canadienne de statistique, on nous a soumis deux ensembles de données.

L'ensemble 1 contient les données de 13 sessions d'évaluation au cours desquelles 91 bières ont été évaluées. Pour chaque bière, nous connaissons la cote moyenne d'appréciation qu'elle a obtenue et une batterie de 35 variables qui la caractérise: 18 variables analytiques telles l'âge, le degré d'alcool, le pH et la couleur entre autres et 17 variables mesurant la concentration d'autant de composés volatils.

L'ensemble 2 contient le même genre d'information quant à 4 autres sessions d'évaluation portant sur 37 bières, sauf que la cote moyenne d'appréciation ne nous est pas transmise.

Nous avons construit, à partir de l'ensemble 1, un modèle de prévision pour la cote moyenne d'appréciation et nous l'avons utilisé pour obtenir des intervalles de prévision de niveau 90% pour les cotes moyennes d'appréciation des bières de l'ensemble 2.

Comme cette analyse a déjà fait l'objet d'un article (Carmichael, Daigle et Rivest,1990), nous nous contentons de la présenter de façon schématique dans la section 2. Par contre, la section 3 contient de la matière nouvelle.Nous y comparons nos prévisions aux vraies cotes (elles nous ont été dévoilées quelques mois après notre analyse) et notre analyse à une analyse automatique.

Enfin,soulignons qu'un deuxième groupe de statisticiens a analysé les mêmes ensembles de données. Leur travail paraît en même temps que le nôtre. Si nous n'y faisons jamais référence,c'est que nous ignorons totalement ce qu'ils ont fait. Ceci devrait vous inciter à aller voir pour vous-même.

2. MODELISATION ET PREVISION

2.1 Etude exploratoire

Nous avons utilisé l'analyse en composantes principales (ACP) robuste (Rivest et Plante 1988) pour étudier les relations entre les variables chimiques et le lissage des nuages de points par la procédure Lowess (Cleveland 1979) du progiciel S (Becker et Chambers 1984) pour étudier la relation entre la cote moyenne d'appréciation et les variables chimiques.

Cela nous a permis de découvrir que l'ensemble 2 était passablement différent de l'ensemble 1 en ce qu'il contenait beau-

coup plus de bières légères (degré d'alcool < 4.5%). Nous en avons tenu compte en décidant d'élaborer deux modèles de prévision: un pour les bières légères et un pour les bières ordinaires. Nous avons été amenés aussi à transformer certaines variables et à en éliminer d'autres: la variable âge, peu liée à la cote, contenant beaucoup d'observations manquantes et un groupe de variables colinéaires au degré d'alcool.

Finalement, nous avons détecté la présence de données délinquantes autant pour la cote d'appréciation que pour certaines des variables chimiques. Nous avons adopté différentes stratégies, à l'étape de la modélisation, pour nous prémunir contre les effets néfastes de ces données.

2.2 Modélisation des bières légères

Pour choisir un modèle de prévision, nous avons utilisé la méthode STEPWISE de la procédure REG de SAS (1985). Après élimination de deux valeurs extrêmes, nous avons obtenu, comme meilleur modèle, une régression linéaire simple de la cote moyenne d'appréciation sur la variable SO2, qui mesure la quantité de sulfites dans la bière.

2.3 Modélisation des bières ordinaires

Nous avons mis plus d'effort sur la modélisation des bières ordinaires, car l'ensemble 1 était beaucoup plus riche en ce type de bières (71 bières sur 91). Un premier groupe de variables pertinentes a été identifié par la méthode STEPWISE. Nous lui avons ajouté un certain nombre de variables, après étude des nuages de points lissés. La procédure RSQUARE a calculé toutes les régres-

sions possibles pour ces variables. Nous avons retenu celle contenant le plus petit nombre de variables et pour laquelle le coefficient C_p de Mallows était plus petit que ce nombre de variables.

Nous avons détecté l'effet de 3 valeurs influentes, en vérifiant la stabilité des estimations obtenues sous 4 critères différents: les moindres carrés, la Proposition 2 de Huber(1981), la minimisation de la somme des valeurs absolues des résidus et la minimisation de la médiane des carrés des résidus.

Comme les procédures de sélection de modèle sont sensibles aux données influentes, nous avons repris l'étape de sélection en l'absence des 3 valeurs influentes pour aboutir en fin de compte au même modèle.

2.4 Intervalles de prévision

Les prévisions des cotes moyennes d'appréciation des bières de l'ensemble 2 sont calculées à partir des deux modèles déterminés auparavant. Pour une bière légère, nous avons dû substituer à la valeur manquante de la variable SO2 la valeur moyenne de cette variable pour les bières légères de l'ensemble 2.

Nous avons construit les intervalles de prévision de manière non-paramétrique, en supposant que les deux modèles donnaient des prévisions de même précision. Les intervalles sont donc de la forme $\hat{y}_i \pm \hat{p}_{90}$, où $\hat{y}_i$ est la cote moyenne d'appréciation prédite sous le modèle approprié, et $\hat{p}_{90}$ est une estimation non-paramétrique du 90e percentile de la distribution de la valeur absolue des erreurs de prévision. Ces erreurs de prévision sont estimées par les résidus PRESS.

3 COMPARAISONS

Voyons premièrement si une modélisation automatique, appliquée sans réflexion, donne de bons résultats.Il s´agit ici d´appliquer la méthode STEPWISE à l´ensemble 1, sans distinguer les bières légères des bières ordinaires et sans se soucier des valeurs influentes. Le tableau 1 présente les coefficients de notre modèle et ceux du modèle automatique.

La figure 1 montre que les intervalles de prévision du modèle automatique sont beaucoup plus longs que ceux que nous avons obtenus. La figure 2 montre que les erreurs de prévision (les différences, dans l´ensemble 2, entre les vraies cotes et les cotes prédites) sont plus petites, en valeur absolue, pour notre modèle(code 1) que pour le modèle automatique(code 2). Donc, il y a beaucoup à gagner à intervenir dans l´analyse comme nous l´avons fait.

Finalement, il est intéressant de comparer la distribution empirique des résidus PRESS, qui estime celle des erreurs de prévision et qui nous a fourni $\hat{p}_{90}$, à la distribution empirique des erreurs de prévision observées sur l´ensemble 2. La figure 3 montre que les erreurs de prévision observées sont plus grandes en valeur absolue que les résidus PRESS. La figure 4 montre le même phénomène de façon différente: la droite indique ce que l´on observe lorsque les deux distributions sont égales. Nous avons découvert que c´est surtout pour les bières ordinaires que les prévisions sont moins bonnes. En fait, il est toujours plus difficile d´obtenir de bonnes prévisions qu´un bon ajustement.

4 REFERENCES

Becker, R.A. et Chambers, J.M. (1984). S: An Interactive Environment for Data Analysis and Graphics. Wadsworth,

Belmont, CA.

Carmichael, J.P., Daigle, G. et Rivest, L.P. (1990). Analyse des préférences des Canadiens en matière de bière. Rev. Can. Statist. 18, 99-108.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. J. Amer. Statist. Assoc., 74, 829-836.

Huber, P.J. (1981). Robust Statistics. Wiley, New York, NY.

Rivest,L.P. et Plante, N. (1988). L´analyse en composantes principales robuste. Rev. Statist. Appl. 36,55-66.

SAS (1985). SAS User´s Guide: Statistics. SAS Institute Inc., Cary, NC.

TABLEAU 1: COMPARAISON DES ESTIMATIONS OBTENUES

NOTRE MODELE

BIERES LEGERES

Variable	Coefficient estimé	Ecart- type
Ord. à l'origine	5.954	0.051
SO2	-0.057	0.019

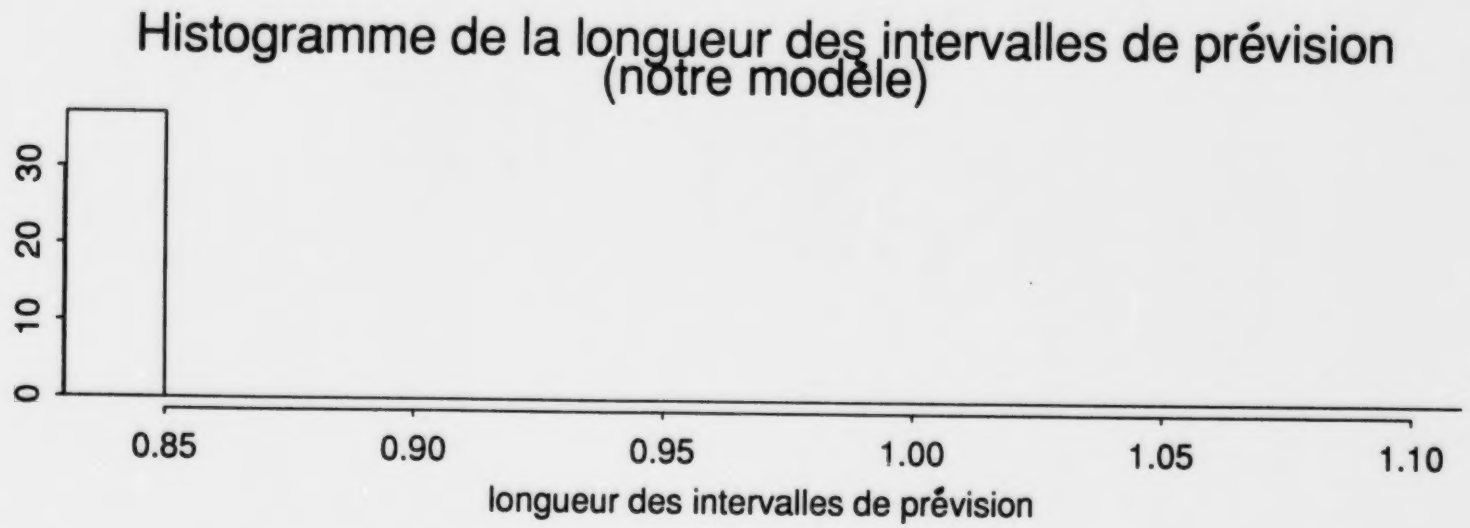
BIERES ORDINAIRES

Variable	Coefficient estimé	Ecart-type
Ord. à l'origine	5.465	0.332
BU	-0.062	0.018
COLOUR	0.337	0.071
ABS(ALCOOL)	2.034	0.399
LOG(FTU)	-0 159	0 063
ISO-BUTANOL	-0 017	0.007
2-METHYL-1-BUTANOL	0.045	0.012
ETHYL DECANOATE	8.022	2.756

MODELE AUTOMATIQUE

TOUTES LES BIERES

Variable	Coefficient estimé	Ecart-type
Ord. à l'origine	8.226	1.28
BU	-0.056	0.016
CO2	-0.804	0.479
RFE	1.592	0.86
ISOBUTYRALDEHYDE	-0.013	0.008
ETHYL PROPIONATE	0 597	1 056
2-METHYL-1-BUTANOL	0.028	0.012
ETHYL DECANOATE	5.86	3.399


Histogramme de la longueur des intervalles de prévision
(notre modèle)
0
10
20
30
0.85
0.90
0.95
1.00
1.05
1.10
longueur des intervalles de prévision

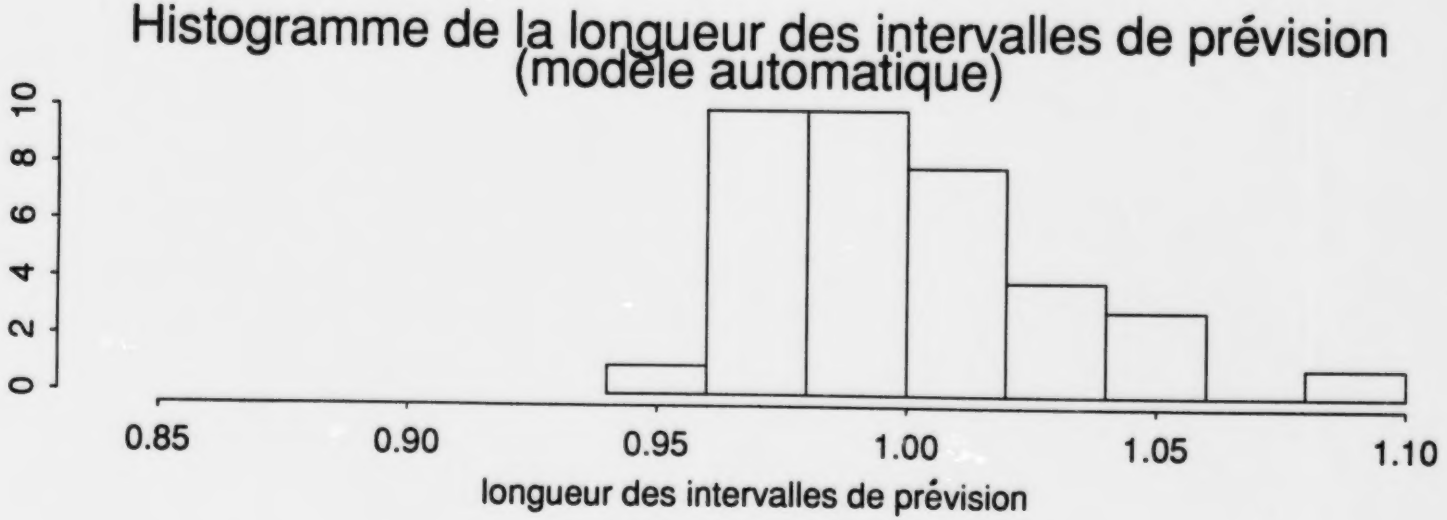

Histogramme de la longueur des intervalles de prévision
(modèle automatique)
0
2
4
6
8
10
0.85
0.90
0.95
1.00
1.05
1.10
longueur des intervalles de prévision

Figure 1

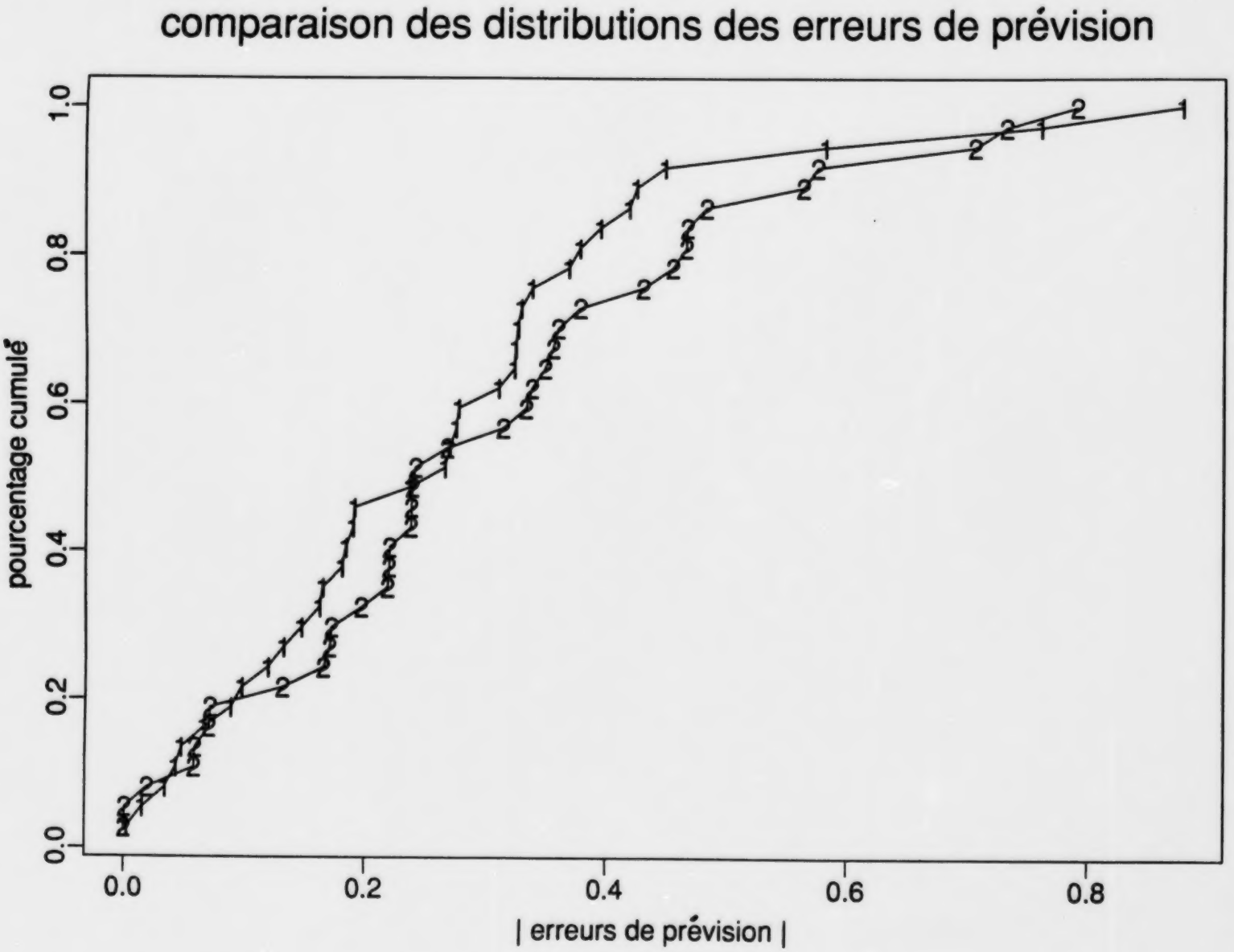

comparaison des distributions des erreurs de prévision
pourcentage cumulé
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
| erreurs de prévision |

Figure 2

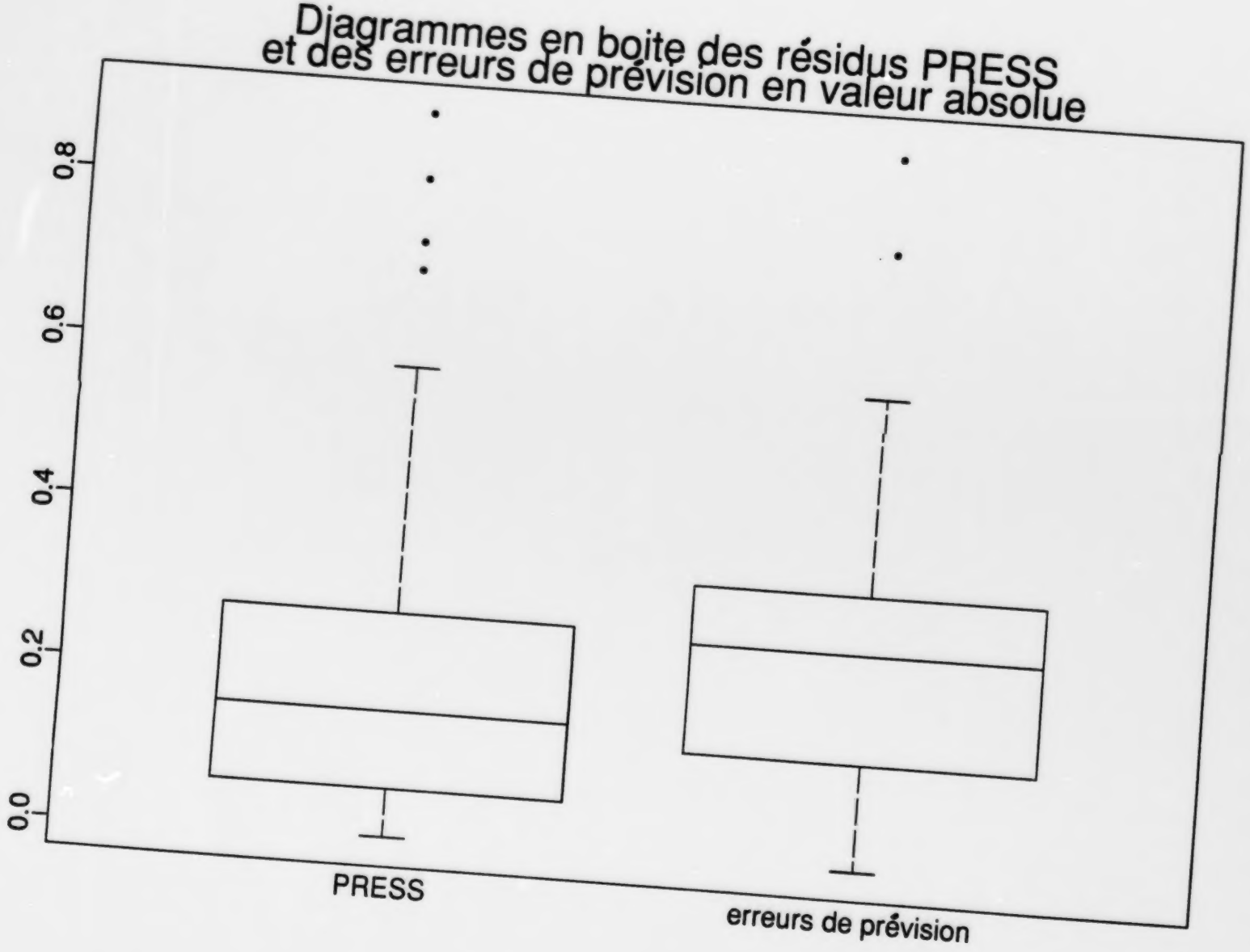

Diagrammes en boite des résidus PRESS
et des erreurs de prévision en valeur absolue
0.0
0.2
0.4
0.6
0.8
PRESS
erreurs de prévision

Figure 3

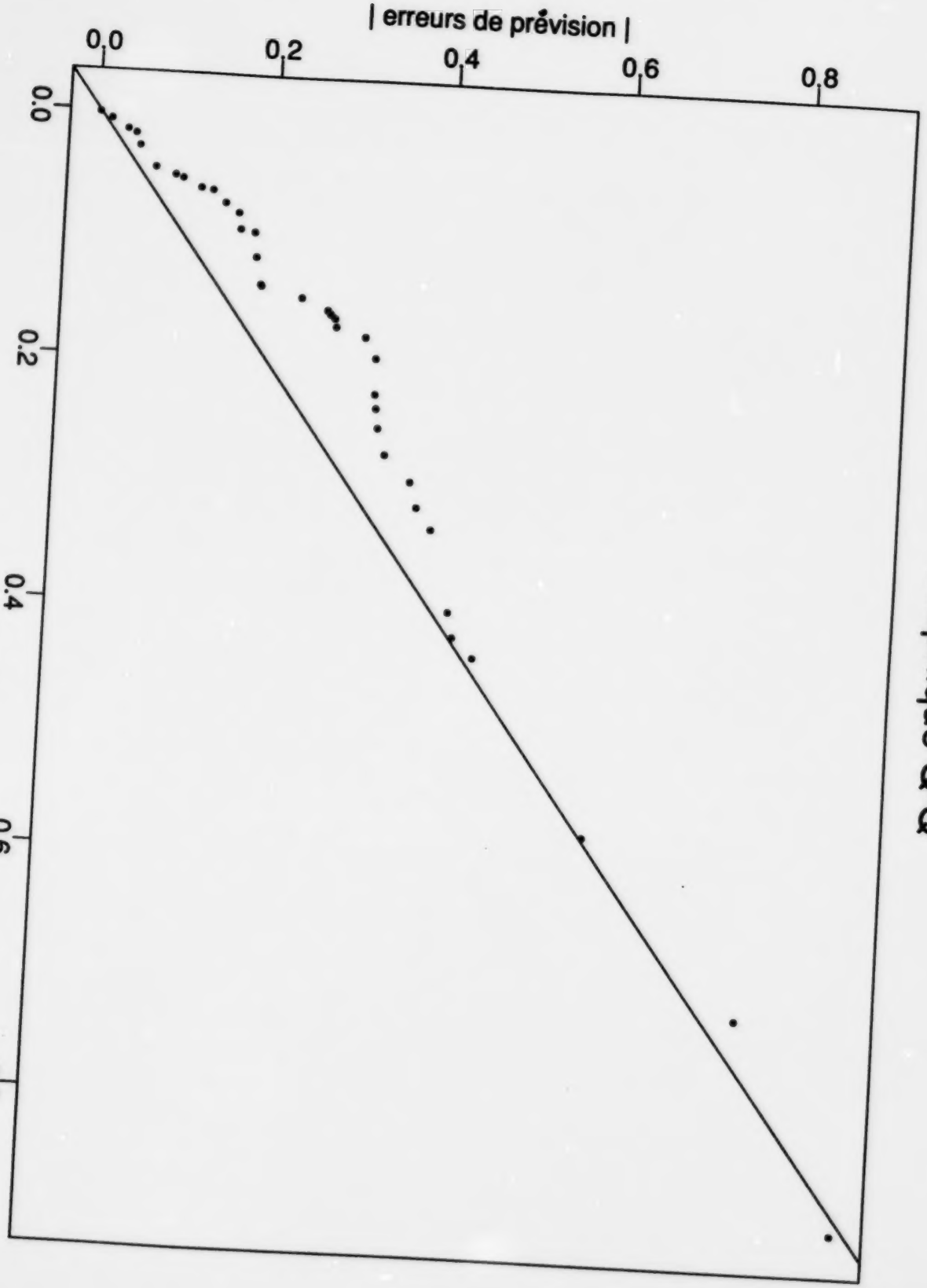

Graphique Q-Q
| erreurs de prévision |
0.0
0.2
0.4
0.6
0.8
| PRESS |
0.0
0.2
0.4
0.6
0.8

Figure 4

Dépouillement d'une enquête statistique relative aux attitudes des infirmières de la région montréalaise

par Bernard Colin (1)
Brenda Mac-Gibbon-T (2)
Jules Tibeiro (3)

INTRODUCTION

Les femmes, dans la société moderne sont généralement "aiguillées" vers des travaux à faibles revenus, objet du préjugé initial dérogatoire à la personne et de disparité des salaires entre hommes et femmes sur le marché de travail en dépit des exigences similaires en matière de formation professionnelle.

La compétence individuelle (scientifique) souffre de plus en plus de ces disparités (inégalités salariales, ingratitude de l'emploi) et les employeurs devraient en être conscients. Malgré les efforts réalisés jusqu'ici par les pouvoirs publics pour attenuer cette discrimination, l'échelle de traitement réservée aux postes occupés par des femmes reste toujours désavantageuse ...

Dans cette étude, nous nous intéressons particulièrement à une carrière en péril, celle des infirmières qui constitue du reste le groupe féminin le plus important au Québec. La vague de protestation s'avère pour le moment un sujet d'actualité préoccupant car intimement lié à l'avenir même de l'appareil sanitaire du Québec.

I. DES DONNÉES AU TABLEAU GÉNÉRAL

L'enquête dont nous rendons compte ici les grandes lignes a été effectuée dans une dizaine d'hôpitaux de la région de Montréal. Le questionnaire (177 questions) a été soumis au départ à 1148 individus dont 53 personnes ayant l'intention de quitter leur poste de travail.

L'ensemble des réponses se prête d'emblée à un codage disjonctif complet vu qu'on impose à tous les individus de choisir, pour chaque question, une réponse et une seule parmi celles qui sont proposées.

Ce tableau croise donc l'ensemble des individus de la population et l'ensemble des réponses à toutes les questions. Il comporte un 1 au croisement de la ième ligne et de la jème colonne si l'individu a choisi la réponse j et des zéros partout ailleurs.

L'efficacité sur le plan descriptif et exploratoire du traitement de ce type de tableau par l'A.F.C.M. (Analyse des Correspondances Multiples) n'est plus à démontrer. Il faut tout de même préciser que cette forme d'analyse s'inspire directement du corpus des données et non des hypothèses de base!

I.1 Principe de la méthode

Notons n le nombre d'individus ayant répondu au questionnaire. On peut associer à chaque question un sous-espace de R^n des fonctions numériques définies sur l'ensemble des individus: le s.e.v. engendré par les variables indicatrices de ses réponses (*les colonnes du tableau disjonctif complet*).

(1) Prof. à l'Université de Sherbrooke, P.Q. - Canada
(2) Prof. à l'Université du Québec à Montréal et à l'Univ. Concordia, P.Q., Canada
(3) Prof. à l'Université de Moncton, N.B. - Canada

Ce sous-espace a pour dimension le nombre des réponses proposées (les variables indicatrices des réponses à une même question étant orthogonales entre elles). C'est donc le sous-espace des fonctions numériques ayant la même valeur pour les individus ayant choisi la même réponse.

Les facteurs de l'A.F.C. peuvent être définis comme les facteurs de l'analyse multicanonique de ces sous-espaces; i.e. le premier facteur est le vecteur de R^n qui rend maximum la somme des cosinus carrés des angles entre ce vecteur et chacun des sous-espaces, le second est orthogonal au premier et maximise le même critère etc...

Tous les sous-espaces considérés contenant la droite des fonctions constantes, le premier facteur est le facteur constant (facteur trivial); les autres facteurs sont orthogonaux à ce facteur constant et sont donc des fonctions centrées (à ce point de vue il y a cohérence entre l'A.F.C. Multiples et l'A.C.P. (Analyse en composantes principales) normée: analyse multicanonique).

Ce point de vue permet ainsi de proposer une projection des questions elles-mêmes sur les facteurs ce qui facilite l'interprétation.

I.2 Le questionnaire

Il comporte, outre le signalétique, 6 volets se rapportant aux 6 thèmes principaux à savoir:

1er volet: l'appréciation générale sur le travail proprement dit

2ème volet: la mise en valeur du répondant

3ème volet: les relations (des infirmières), perception de management

4ème volet: les capacités, chances de promotion (perception de la carrière)

5ème volet: le jugement sur l'Institution (facteur de loyauté)

6ème volet: les contraintes liées aux réglèments de l'hôpital, au travail et les contraintes personnelles)

II. TABLEAUX ANALYSÉS ET RÉSULTATS PRÉLIMINAIRES

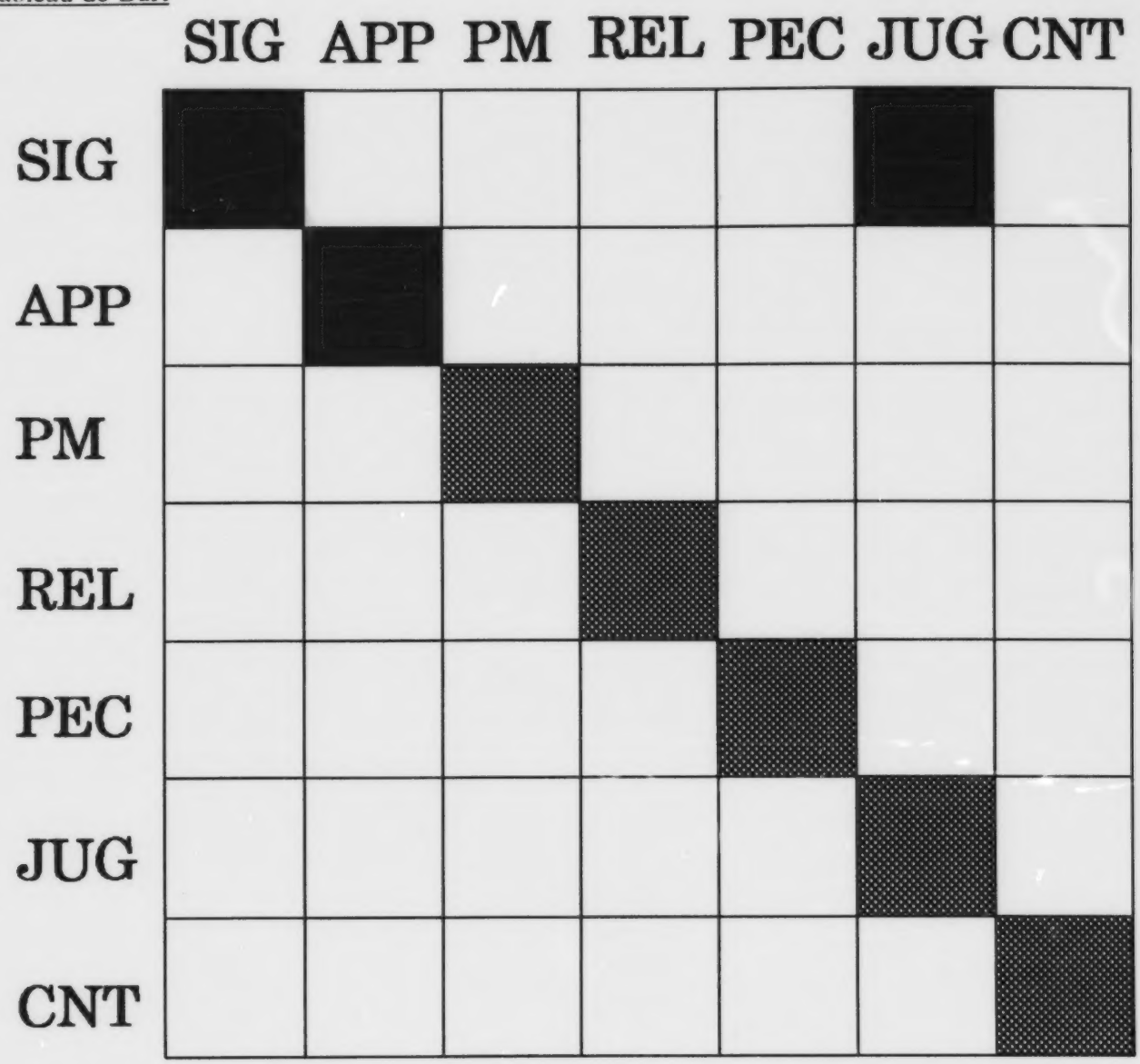
II.1 Le tableau des réponses individuelles

Les 162 questions (dont on a épuré des questions redondantes générant ipso facto des non-réponses) sont éclatées en 507 modalités. D'où le tableau binaire kIJ à 1148 lignes et 507 colonnes. Le codage linéaire par morceaux (Laboratoire de Paris VI, Pr. J.P. Benzécri) s'effectuera ultérieurement sur le tableau kIJ 660X507 dans le cadre d'une étude comparative eu égard aux turbulences occasionnées par les inévitables données manquantes dans certains thèmes de l'enquête.

II.2 Le tableau de Burt

C'est le tableau carré J x J (507 x 507); avec B(j,j') = nombre de sujets ayant simultanément adopté les modalités j et j'.

Tableau de Burt



SIG: Signalétique
APP: Appréciation générale sur le travail
PM : Perception du Management
REL: Relation des infirmières
PEC: Perception de la carrière (chances de promotion)
JUG: Jugement sur l'Institution(l'hôpital)
CNT: Contraintes liées aux règlements de l'hôpital

$K'_{jj} = K_{ij}' \times K_{ij}$ où K_{ij}: tableau des réponses individuelles

$k'(j,j') = \Sigma\{ k(i,j) \times k(i,j') : i \epsilon I \}$

$= \text{Card} \{ i : i \epsilon I ; k(i,j') = 1 \}$

On trouve donc à la croisée de la ligne j et de la colonne j' le nombre k'(j,j') des individus ayant simultanément adopté les modalités j et j'

■ : tableau retenu

II.3 Les tableaux retenus

SIGNALETIQUE * SIGNALETIQUE

JUGEMENT * JUGEMENT

SIGNALETIQUE * JUGEMENT

III. LES RÉSULTATS

III.1 Sur le signalétique

Vue d'ensemble des résultats

40% d'inertie sur les 7 premiers axes.

Le 1er axe (*axe d'ancienneté au travail*) explique près de 12% de l'inertie totale du nuage. Il oppose essentiellement les jeunes infirmières - âgées de moins de 30 ans, célibataires, jeunes diplômées, au poste à partir de 1981 - aux autres (âgées de plus de 30 ans, mariées, au moins un enfant et excercant leur emploi à partir de 1941 ...)

Il faut signaler le fait que les plus jeunes sont à leur poste de service généralement depuis moins de 2 ans, ont eu plus de 2 jobs durant les 5 dernières années et travaillent surtout dans les services médicaux (et de chirurgie)

Les plus âgées, ayant, nous imaginons, plus d'expérience, oeuvrent particulièrement dans des services spécialisés notamment la gynécologie (Obstétrique) et dans des salles d'opération.

Le 2ème axe explique près de la moitié de l'inertie associée au 1er axe. Il est fait de l'opposition entre le *niveau d'études* Baccalauréat * Diplôme en Nursing. Cet axe révèle aussi une opposition à un moindre degré entre la clinique ambulatoire, la psychiatrie et les autres postes médicaux.

III.2 Facteur de loyauté (Jugement sur l'Institution)

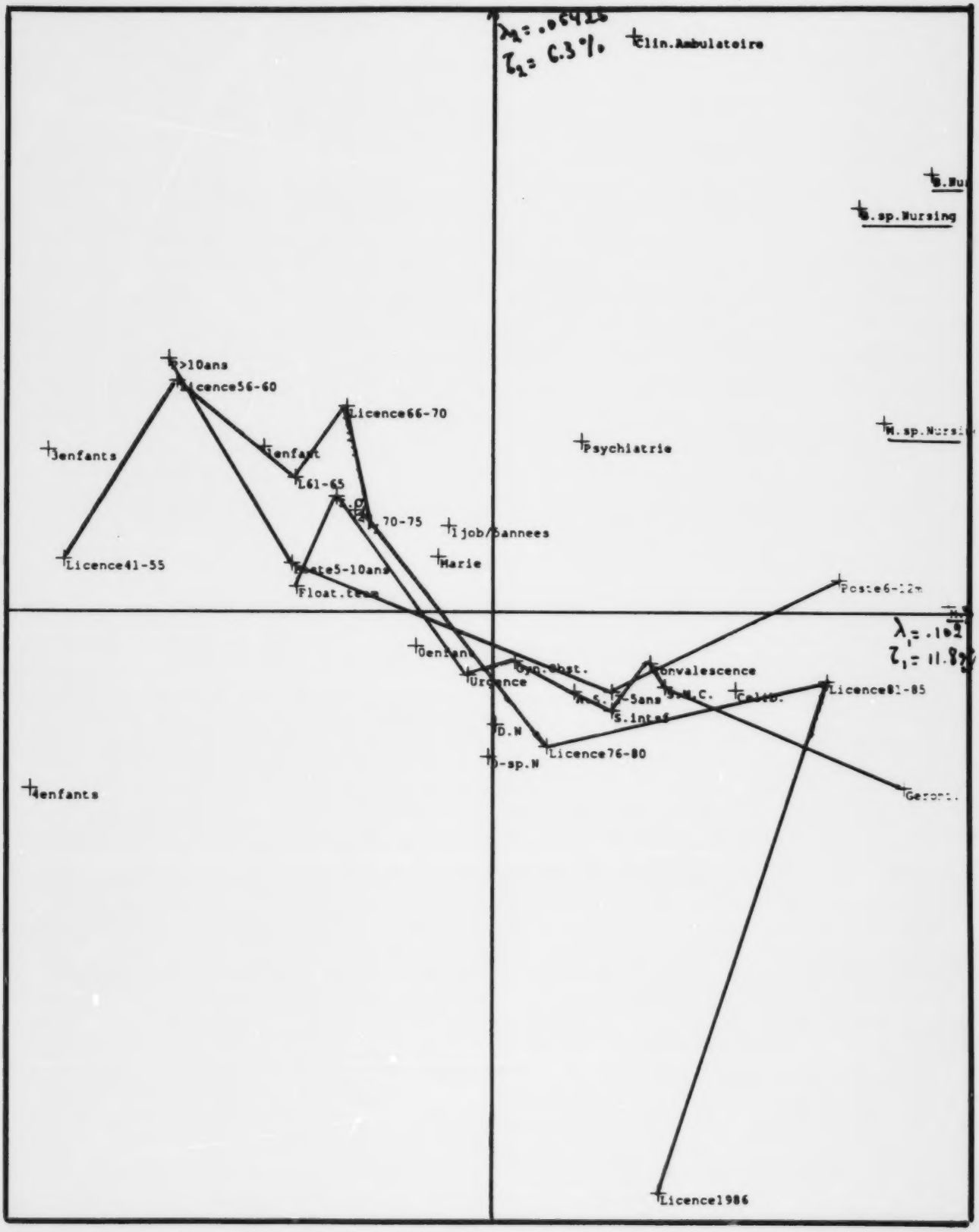
68% de l'inertie sur les 7 premiers axes.

Le 1er axe explique à lui seul 36.5% de l'inertie totale. Ce facteur, nettement détaché du deuxième (T = 16.4%) représente un peu plus de la moitié de l'inertie totale du nuage: c'est un *facteur de niveau général de jugement sur l'institution de travail.*

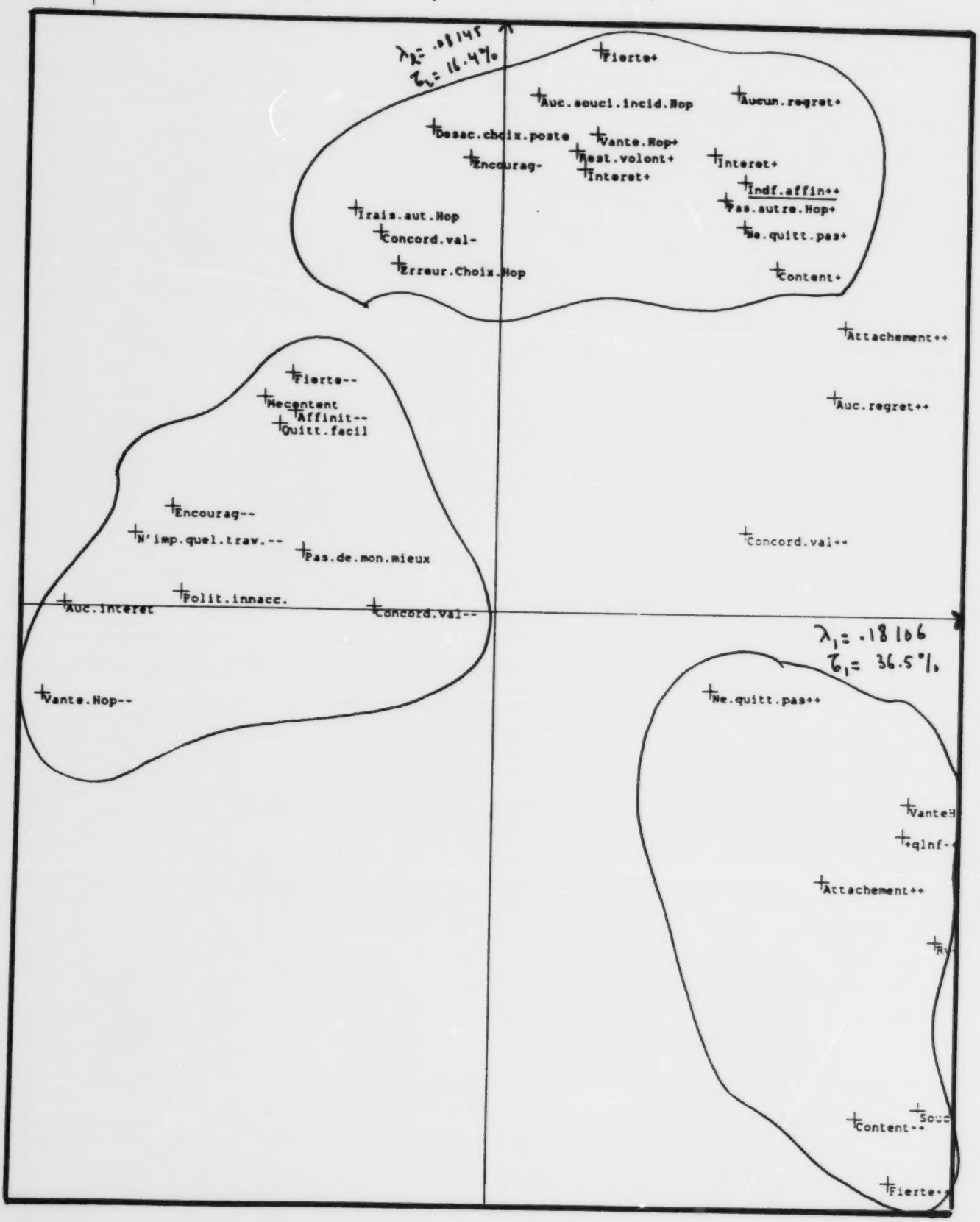
Du côté F1 > 0, les attitudes associées sont en désaccord sur les questions proposées lesquelles sont présentées sous forme d'affirmations.

Le plan (1 x 2) apparaît dominé par 4 regroupements de modalités d'opposition (Accord total, désaccord total, accord et désaccord *sur le bout des lèvres*). Ces 2 derniers regroupements font justement l'opposition qu'on retrouve sur le 2ème axe.

ligne 1 + ligne 2



facteur de loyauté + facteur de loyauté



Ces modalités d'opposition sont des facteurs d'enrichissement (d'évolution) comme le degré de défi personnel, la variété des tâches, la prise de responsabilité, la perfection des services, l'implication personnelle et l'ajustement entre l'habileté, l'intérêt et le travail proprement dit.

III.3 Signalétique * Facteur de loyauté

62% d'inertie sur les 7 premiers axes.

Sur le 1er axe, les 'individus' qui contribuent le plus à sa création sont:
du côté négatif:

I9: Mère de 3 enfants
I7: Mère de 2 enfants
I25: au service depuis plus de 10 ans
I26: 1 job durant les 5 dernières années
I50: St-Mary's Hospital et particulièrement I45 (l'hôpital des enfants malades) dont la contribution est la plus importante (QLT = 884, COR = 704 et un très fort facteur négatif).

A ces "individus" s'associent sur le même demi-plan négatif les modalités (relatives à certains stéréotypes de jugement) ci-après:

I334: "Je m'en fous complètement"
I344: "Je n'ai aucune affinité avec cet hôpital"
I349: "Je ne ressens aucune fierté d'appartenir à cet hôpital"
I359: "Je suis très mécontent d'avoir choisi de travailler dans cet hôpital"

du côté positif

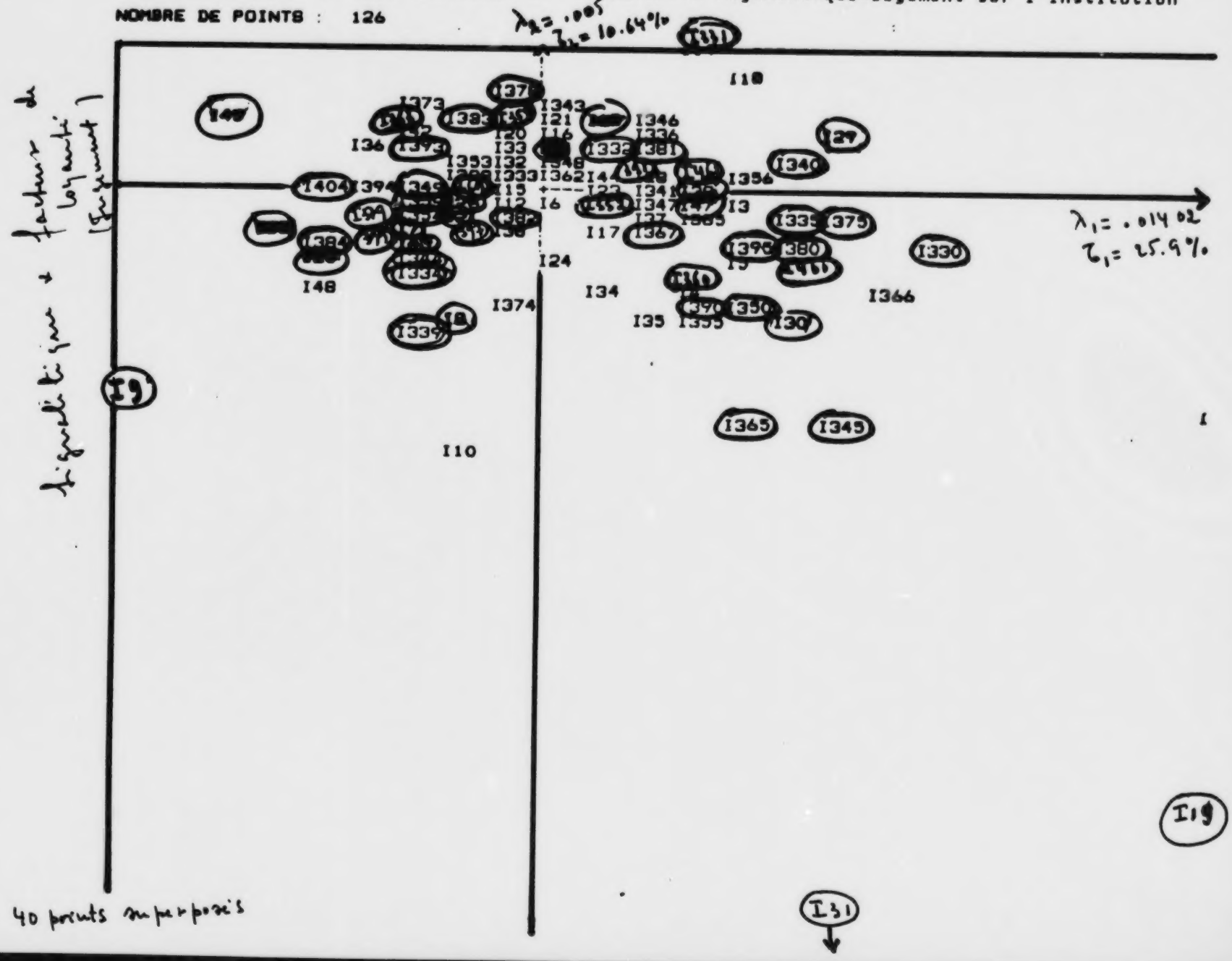
les modalités relatives au signalétique sont:

I19: Niveau de Maîtrise (autre)
I29: 4 jobs durant les 5 dernières années
I39: Soins intensifs
I47: Hôpital Général de Montréal

auxquelles s'associent les modalités de jugement suivantes:

I330: "Je me soucie beaucoup sur les incidents qui peuvent survenir à l'hôpital"
I335: "Je fais plus qu'il n'en faut"
I350: "Je ne quitterais jamais cet hôpital"
I380: "Il y a concordances des valeurs entre celles de l'hôpital et les miennes
I385: "Je ne pourrai pas aller travailler ailleurs"
1400: "Je travaille dans un très bon hôpital"

Il ressort de cette analyse le fait que les plus qualifiés au niveau académique (les plus jeunes) sont généralement satisfaits de leur travail et cultivent un certain attachement à leur hôpital contrairement aux plus expérimentés qui remettent plus en cause le système (milieu) de travail que la profession proprement dite ce qui explique tout de même leur ancienneté.


AXE HORIZONTAL(1)--- AXE VERTICAL(2)---TITRE: A.F.C Signalétique*Jugement sur l'institution
NOMBRE DE POINTS : 126
λ2= .005
τ2= 10.64%
λ1= .014 02
τ1= 25.9%
Signalétique + facteurs du Jugement
40 points superposés
I18
I24
I34
I10
I374
I48
I35
I366
I117
I124
I131
I365
I345
I330
I340
I356
I335
I375
I380
I398
I404
I147
I19
I31

IV. Conclusion

On peut conclure, à la lumière de cette étude - qui commence! - que le mécontentement des infirmière(e)s en milieu de travail provient principalement d'un travail perçu comme trop accaparant (exigeant) allié à une perception négative de l'encadrement et à un degré beaucoup moindre des salaires.

Contrairement, à ce que l'on pourrait penser, il ne semble pas que l'échelle des salaires soit une source profonde d'insatisfaction. On peut remarquer par ailleurs comme on est en droit de s'y attendre que le dynamisme et l'esprit d'initiative se rencontrent surtout chez les groupes de jeunes infirmières (24 à 30 ans, célibataires et mariées sans enfants) pour laisser place parmi les plus âgées d'entre elles à une certaine indifférence et perception négative du travail dues principalement à leur mise en valeur (thème qui sera étudié dans la suite de ce travail particulièrement l'impact des secteurs de travail dans les hopitaux sur le comportement des infirmières).

Bibiliographie

(1) J.P. Benzécri et coll. - L'analyse des données, Tome 1, Tome 2, Dunod, 1973.

(2) J.P. et F. Benzécri - Analyse des correspondances: exposé élémentaire, Dunod, 1980.

(3) M. Volle - "Une méthode pour lire et commenter automatiquement de grands tableaux statistiques", Economie et Statistique, no. 52, janvier 1974.

(4) C. Bastin, J.P. Benzécri, C. Bourgarit, P. Cazes - Analyse des données: Abrégé théorique; études de cas modèles, Dunod, 1980.

(5) R. Amalberti, J.P. Benzécri, N. Tabard - Techniques de la description statistique, Dunod, 1977.

(7) D.G. Dialetis, M.G. Meimaris - "Elaboration d'une typologie de la population d'Athènes fondée sur les problèmes de la ville et le profil du maire désiré" in Data Analysis and Informatics, E. Diday et. al., Ed. North Holland Publishing Company, 1980.

(8) Abou-Fahrat - "Comportement et jugements vis-à-vis de l'usage du tabac au liban" in C.A.D., vol. VIII, no 2, 1983 - pp. 133-158.

UTILISATION DE TECHNIQUES D'ANALYSES FACTORIELLES ET DE CLASSIFICATION AUTOMATIQUE DANS LE TRAITEMENT D'UN CORPUS DE DONNEES SOCIOLOGIQUES

Marie-Marthe Cousineau et Guy Cucumel[1]

Département de sociologie, UQAM,
CP 8888, Succ. A, Montréal (Québec)
H3C 3P8

Introduction

Sans vouloir procéder à l'historique du développement des méthodes d'analyse des données, rappelons simplement que le traitement de gros corpus à l'aide de techniques complexes ne s'est développé qu'au milieu des années 1960 et a pris un essort considérable au cours des cinq dernières années. La mise à la disposition des chercheurs et praticiens de logiciels et progiciels de plus en plus nombreux et conviviaux, aussi bien sur ordinateurs centraux que sur micro-ordinateurs, a largement contribué à changer les pratiques de recherche dans tous les domaines, mais encore plus particulièrement en sciences humaines et sociales. Alors que jusqu'à une époque récente, il se contentait de réaliser des analyses univariées (fréquences, calculs de moyennes, d'écarts-types...) et des croisements bi ou - pour les études les plus audacieuses - multivariées, le chercheur en sciences sociales s'intéresse de plus en plus à des traitements statistiques complexes. Pour notre part, nous avons choisi de soumettre à des analyses plus poussées un corpus de données initialement traitées à partir d'une série d'analyses uni et bi-variées.

Les résultats de ce deuxième traitement des données allaient remplir deux objectifs: d'une part permettre de mieux circonscrire les caractéristiques de la population à l'étude - essentiellement en termes de profil-type -, d'autre part, montrer comment une méthodologie s'appuyant sur des méthodes multidimensionnelles d'analyse des données permet de vérifier ou d'infirmer des intuitions, des hypothèses issues d'analyses classiques et de mieux saisir la complexité de la réalité. La recherche de profils-types précisément nous a conduit à privilégier l'analyse des correspondances multiples et la classification hiérarchique comme mode de traitement pour cette deuxième phase d'analyse.

1. Un cas d'exemple : les itinérants de Montréal

En 1988, Shirley Roy dressait un profil socio-économique des hommes itinérants montréalais. L'étude de Roy, qui se voulait à l'origine essentiellement descriptive, a permis de brosser un premier portrait de la clientèle fréquentant l'Accueil Bonneau (un centre de dépannage pour itinérants où se retrouvent, à un moment ou à un autre et pour différentes raisons, une bonne partie des itinérants du centre-ville montréalais). Cette étude a aussi permis de suggérer un certain

---

[1] Chercheurs au département de Sociologie de l'UQAM, membres du GRAPPP (Groupe de Recherche et d'Analyse sur les Politiques et les Pratiques Pénales).

nombre d'hypothèses, de pistes d'analyse et de conclusions qui à leurs tours ont permis de poser les bases nécessaires à une lecture renouvelée de la clochardise et de l'itinérance[1].

Les itinérants, on s'en doute, constituent une population qui pose des problèmes d'échantillonnage importants dont nous ne discuterons pas ici. Précisons simplement que les données ont été recueillies à l'aide de questionnaires remplis par les interviewers - pour la plupart des bénévoles oeuvrant à l'Accueil Bonneau - au cours d'entretiens menés auprès de 295 bénéficiaires fréquentant régulièrement cette institution.

Le corpus constitué par Roy (1988) se compose à l'origine de 5 blocs de variables. Le premier bloc est constitué, d'une part, de variables socio-démographiques traditionnelles (âge, état civil, type de résidence occupée, niveau de scolarité, source de revenu) et, d'autre part, de variables plus spécifiques (3 derniers emplois occupés, durée des 3 derniers emplois occupés, stabilité du dernier emploi). Le second bloc est composé de variables relatives à l'itinérance (nombre de déménagements au cours de la dernière année, manque de gîte, nombre de jours sans gîte). Le troisième bloc s'intéresse au passé psychiatrique et carcéral du répondant (séjour en institution psychiatrique, séjour en prison, séjour en pénitencier). Le quatième bloc concerne quant à lui les habitudes de consommation et le principal problème rencontré par les itinérants (fréquence de la consommation alcoolique, fréquence de la consommation de drogue, principal problème : solitude, santé, chômage, ...). Enfin, le cinquième bloc de variables, que nous ne retiendrons pas pour les fins de la présente démonstration, porte sur la fréquence d'utilisation des services par les bénéficiaires.

2. L'intuition de départ

Les données traitées dans un premier temps par Roy lui ont permis d'entrevoir, à partir de la conjonction d'une série de croisements bi-variés, que le groupe des itinérants est hétérogène; il serait composé de deux sous-groupes qui se détachent de manière plus ou moins marquée de l'image que l'on se fait traditionnellement du clochard âgé, mal fagoté, mendiant au coin des rues... Roy, en regard des indices qui lui permettaient de les distinguer, nommait ces deux sous-groupes **les précaires et les errants.**

Le groupe des errants correspondrait à l'image généralement la plus répandue des clochards. S'y retrouvent des solitaires qui n'ont aucun domicile fixe, dorment dans les parcs, les lieux publics, près des bouches de métro et qui ne bénéficient d'aucun revenu. Il s'agirait du groupe le plus visible, mais pas nécessairement celui qui compte le plus d'individus, auquel on a généralement tendance à assimiler l'ensemble des itinérants.

Le groupe des précaires, le plus important numériquement selon les premières constatations de Roy, se composerait, pour sa part, des itinérants qui vivent dans des conditions matérielles minimales. Contrairement aux errants, ceux-ci reçoivent généralement des prestations d'aide sociale et possèdent un lieu de résidence, condition essentielle au fait qu'ils reçoivent des prestations d'aide sociale.

Nous avons repris les données traitées par Roy et nous les avons soumis à une analyse des correspondances multiples et des classifications. Ce que ces analyses montrent, c'est que les pistes avancées par Roy non seulement se vérifient statistiquement mais se précisent et se compléxifient.

---

[1] *Seuls dans la rue*, paru en 1988 aux Editions St-Martin, présente les conclusions de Roy issues de ces premières analyses.

3. La méthodologie

La figure 1 résume l'enchaînement des méthodes que nous avons prévilégié. Les tris à plat, réalisés dans un premier temps permettent une observation descriptive des données et un regroupement éventuel de certaines modalités en fonction des distributions de fréquences. Les croisements bi-variés donnent une idée des liaisons entre variables. L'analyse des correspondances multiples permet d'identifier les grandes tendances de la population à l'étude et en fait ressortir les caractéristiques les plus significatives. Elle apparaît comme un résumé de tous les tableaux croisés des variables dont on dispose.

Une fois réalisés ces premiers traitements, on procède à une classification hiérarchique sur les coordonnées factorielles en utilisant l'indice de Ward afin de repérer des profil-types issus de la population. Cette classification donne une idée du nombre de classes présentes dans la population en coupant l'arbre hiérarchique au niveau d'un saut important de l'indice hiérarchique. La partition est optimisée par réaffectation de certains points à l'aide de l'algorithme des centres mobiles. On interprète la partition par les modalités les plus caractéristiques.

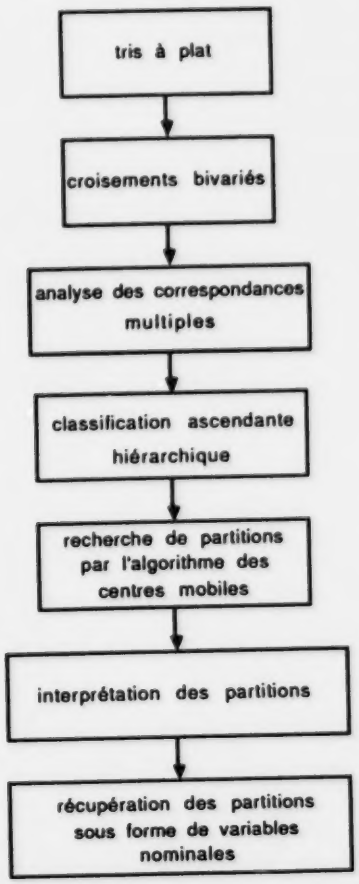


figure 1: enchaînement des méthodes

4. Le choix des variables

Le choix des variables actives et illustratives est toujours extrêmement délicat. Dans des recherches de type causal, les variables actives sont les variables explicatives et les variables illustratives les variables à expliquer. Dans le cas d'une analyse exploratoire, particulièrement lorsqu'on recherche des profils-types, tant les variables de la problématique à l'étude que les variables socio-démographiques peuvent être utilisées comme variables actives ou, au contraire, être projetées en illustratives. Dans la présente étude, les variables du premier et du second blocs ont été utilisées en actives, ainsi que les fréquences de consommation alcoolique et de consommation de drogue. Trois variables que l'on contrôle difficilement ont été projetées en illustratives : le séjour en institution psychiatrique, le séjour en prison et le principal problème dénoncé par les itinérants.

5. Les résultats

5.1 L'ACM

Plutôt que de concentrer sur un seul graphique surchargé l'ensemble des résultats issus de l'analyse factorielle, nous avons choisi de les représenter en trois schémas tirés du même plan factoriel, chacun des schémas rendant compte d'une partie de l'information. La consultation simultanée des plans factoriels 2a, 2b et 2c permettra de constater que de l'analyse des correspondances multiples deux grandes tendances se dégagent.

Considérant d'abord la figure 2a, on constate que sur le premier axe s'opposent, à une extrémité, un premier groupe de clochards qu'on pourrait considérer comme étant les plus stables, du moins en en ce qui a trait aux conditions de logement, c'est-à-dire ceux qui se caractérisent par le fait qu'ils possèdent un lieu de résidence (chambre ou appartement), qu'ils n'ont jamais ou une seule fois déménagé et qu'ils n'ont jamais vraiment manqué de gîte, alors qu'à l'autre extrémité de l'axe se trouvent ceux qui ont dû au cours de la dernière année faire face à plus de deux déménagements et qui ont manqué de gîte pour des périodes allant de 1 à 5 jours à plus de trois mois. On remarquera aussi que les itinérants du premier groupe, qui pour une bonne part reçoivent des prestations d'aide sociale, disent souffrir principalement de problèmes de santé, alors que du côté des moins stables, qui eux ne jouissent d'aucun revenu, on dit souffrir surtout du chômage et du manque d'argent. Enfin, on notera que les premiers, contrairement aux seconds, affirment n'avoir jamais séjourné en institution psychiatrique.

La figure 2b montre pour sa part que, toujours sur le premier axe, les clochards les plus âgés (ceux de 40 ans et plus) et les moins scolarisés (entre 0 et 9 ans de scolarité) qui affirment pour une bonne part ne pas consommer de drogues s'opposent aux plus jeunes (moins de 40 ans) qui apparaissent plus scolarisés et qui avouent consommer des drogues soit de temps en temps soit, pour certains, souvent.

Enfin, la figure 2c témoigne du fait que, encore une fois sur l'axe 1, on trouve du côté de ceux qu'on qualifiaient de plus stables en regard des conditions de logement, ceux qui ont occupé au cours de leur vie au moins une fois un même emploi pour une durée égale ou supérieure à huit ans et, du côté des plus instables, ceux qui tout en occupant des emplois apparemment moins précaires (ouvriers spécialisés et emplois de direction) n'ont toutefois pas bénéficié au dernier moment de leur vie active, d'un emploi de longue durée.

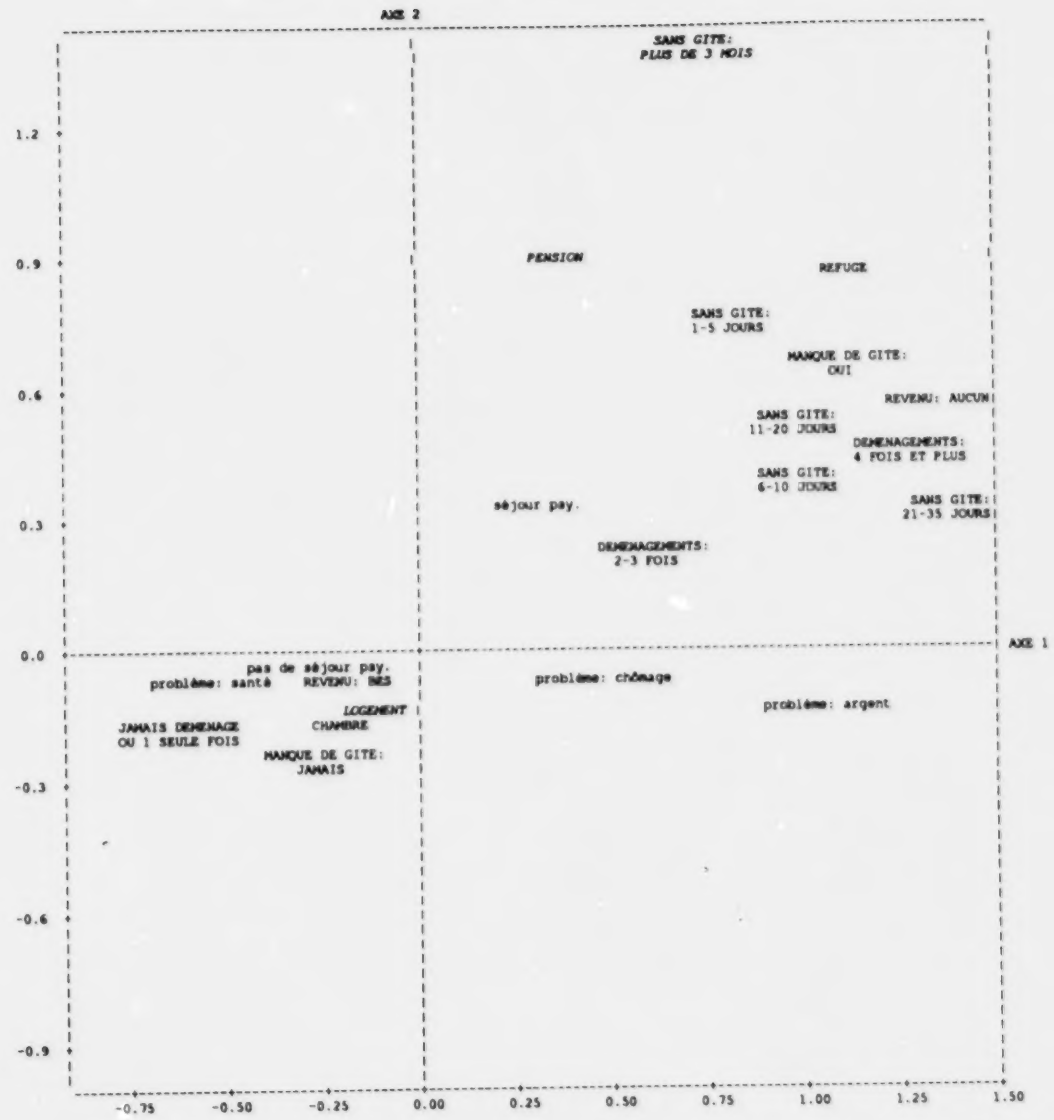


figure 2a: caractérisation des axes par la stabilité de résidence et les principaux problèmes vécus par les itinérants

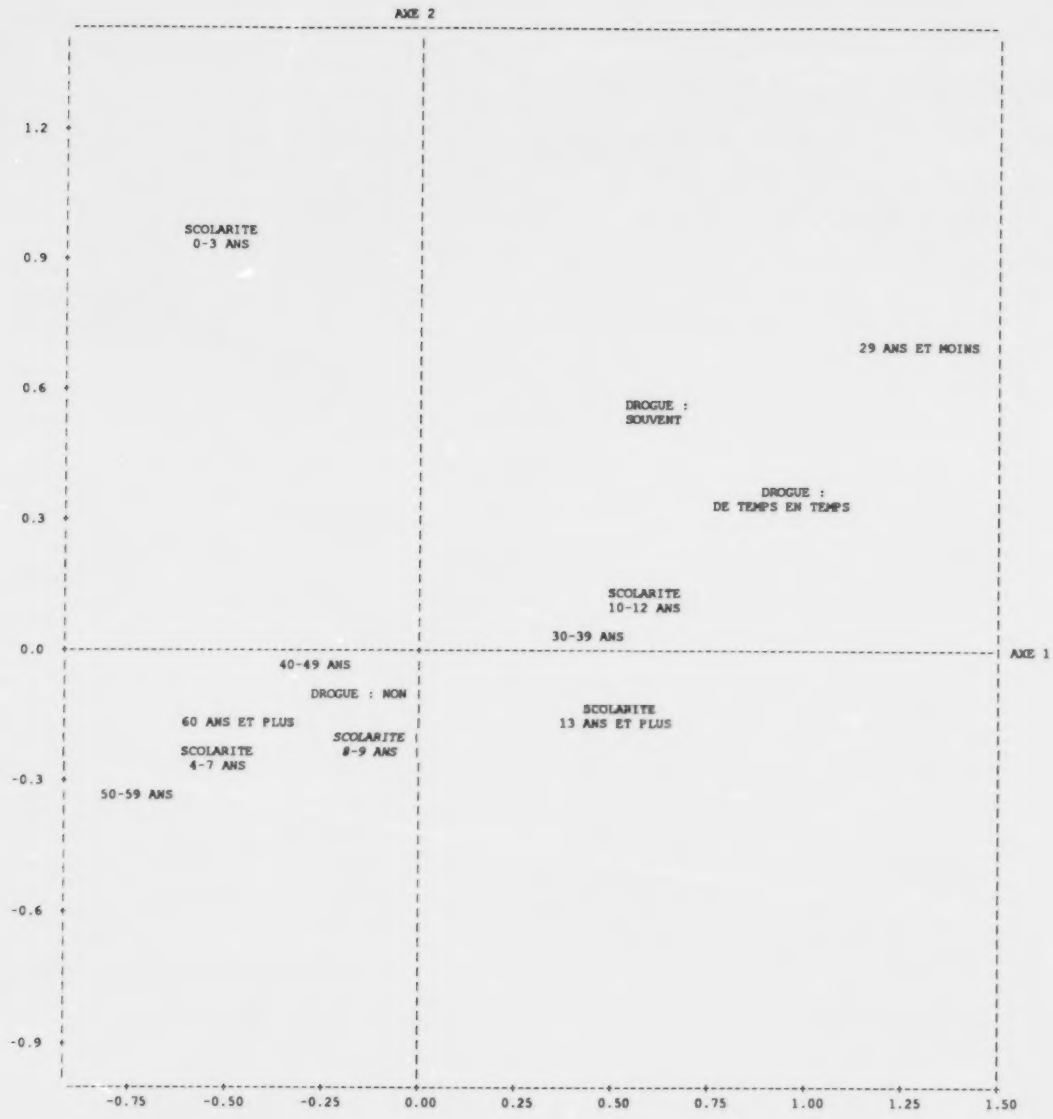


figure 2b : caractérisation des axes par l'âge et la scolarité

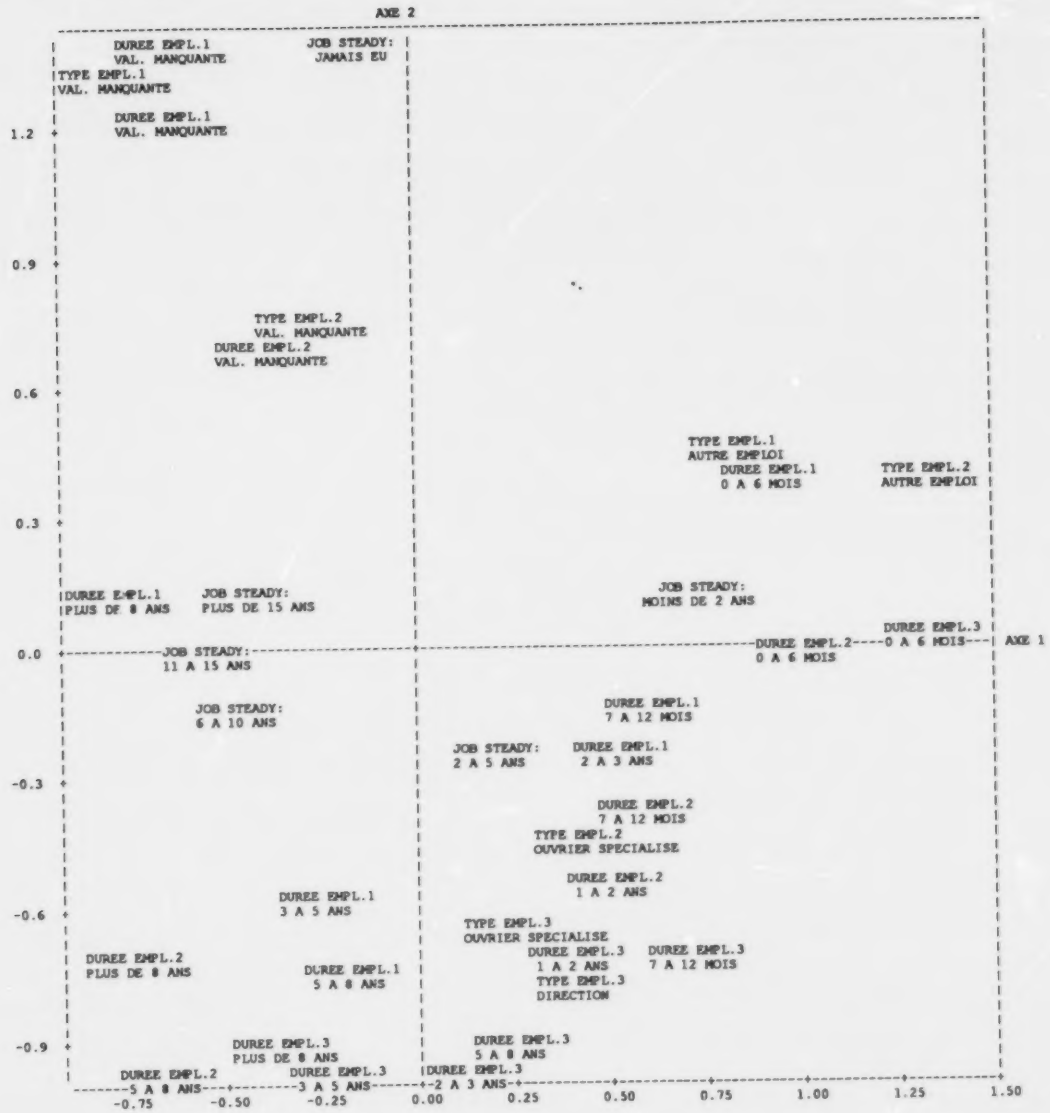


figure 2c: caractérisation des axes par l'emploi[1]

[1] EMPL.1 réfère au dernier emploi occupé, EMPL.2 à l'avant dernier emploi occupé, etc.

Le deuxième axe, pour sa part, est beaucoup moins parlant. Seule la figure 2c rend compte d'une opposition entre, d'un côté de l'axe, les clochards n'ayant jamais eu de "job steady" et, de l'autre, le reste de la population.

L'ACM met donc en évidence des oppositions qui rejoignent celles qu'avaient pressenties Roy entre d'un côté des clochards qui cumulent un ensemble de caractéristiques les rapprochant des véritables errants, qui sont représentés par les jeunes qui ont fait face à des conditions de travail difficiles, qui ont souvent manqué de gîte, qui disent souffrir surtout du chômage et du manque d'argent et avouent consommer de la drogue de temps en temps, et d'un autre côté un autre groupe d'itinérants plus âgés et plus stables, bénéficiant ceux-là de conditions de vie minimales, c'est-à-dire qu'ils possèdent une adresse et reçoivent régulièrement des prestations d'aide sociale. Ces derniers disent souffrir principalement de problèmes de santé. La réalisation d'une classification hiérarchique va permettre de mieux cerner les profils-types présents dans la population à l'étude.

5.2 La classification ascendante hiérarchique

La figure 3 montre la partie supérieure du dendrogramme obtenu par la classification hiérarchique. Seuls les paliers les plus hauts, c'est-à-dire ceux associés à des classes d'effectifs importants sont représentés.

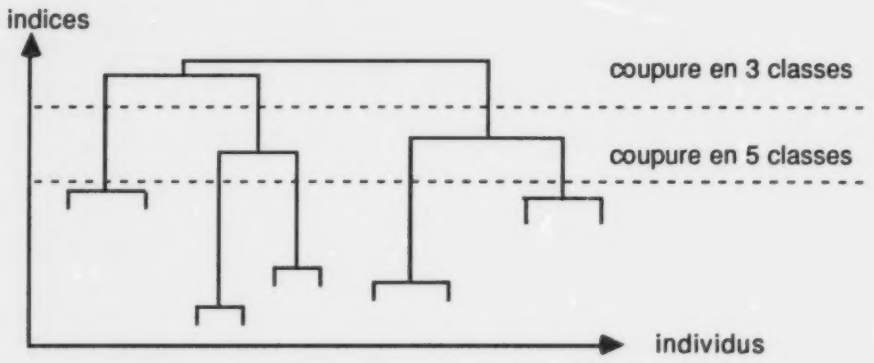


figure 3 : arbre hiérarchique

On constate que l'arbre hiérarchique suggérait, a priori, une classification en trois classes ou en cinq classes. Ces deux coupures correspondent à des sauts importants de l'indice hiérarchique, signifiant que les deux classes réunies avant le saut, bien qu'étant les plus proches, sont relativement éloignées, reflétant ainsi une bonne hétérogénéité des classes entre elles et une bonne homogénéité interne.

Toutefois, afin de mieux cerner les profils-types de la population, en autant que ceux-ci existent, on a procédé successivement à plusieurs classifications impliquant un nombre de plus en plus grand de classes. On étudie donc les coupures en deux, trois, quatre et cinq classes afin de voir de quelle façon s'organisent les groupes à chacune de ces différentes étapes. La figure 4 illustre ce partitionnement. Les nombres en italique indiquent l'évolution des flots lors de la césure des classes. On constate que les effectifs des classes après césure ne correspondent pas nécessairement

exactement aux flots qui les précèdent. Ce fait est dû à la procédure de réaffectation de l'algorithme des centres mobiles qui optimise les partitions obtenues après coupure de la hiérarchie.

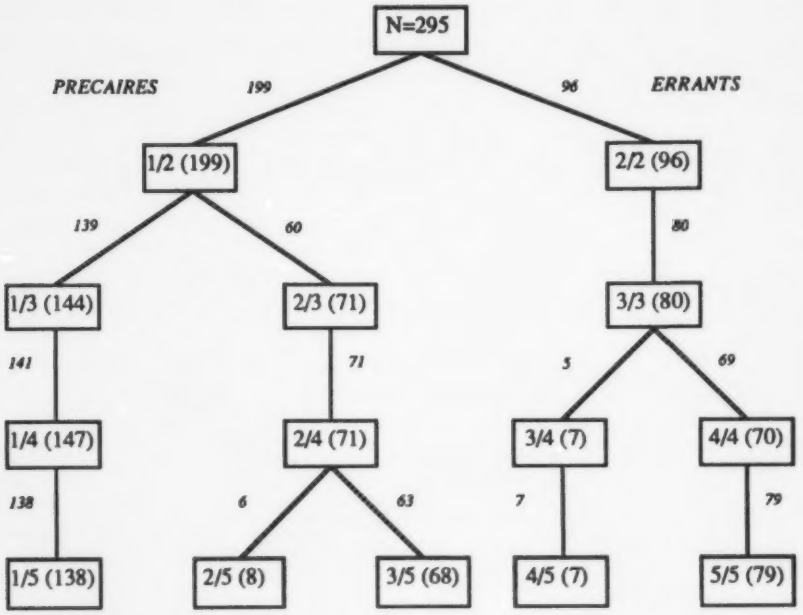


figure 4 : partitionnement en 2, 3, 4 et 5 classes

L'interprétation des partitions se fait de façon simple. Pour chaque classe, sont sélectionnées les modalités significatives au sens de la statistique du chi-carré associée au tableau de contingence croisant la modalité et la classe concernées. Le seuil de sélection correspond à un seuil de rejet de 5% dans une situation de test. Le tableau ci-après présente les caractérisations des classes ainsi obtenues. Seuls sont mentionnées les modalités caratéristiques par ordre décroissant de chi-carrés significatifs. La partie gauche du tableau donne les caractéristiques du partitionnement en deux classes. La colonne de droite indique les caractéristiques des 3 nouvelles classes qui émergent lors des coupures sucessives en 3,4 et 5 classes.

Une coupure en deux classes précise l'intuition de Roy en faisant apparaître un premier groupe qui cumule les caractéristiques suivantes: il s'agit des clochards les plus agés qui possèdent une adresse (ils vivent pour la plupart en chambre), ont peu ou pas déménagé au cours de la dernière année et donc n'ont jamais véritablement manqué de gîte. On peut facilement associer ce premier groupe (classe 1/2) à celui que Roy nommait le groupe des précaires. Il s'agit du groupe numériquement le plus important; s'y retrouvent plus des deux tiers (67,5%) de la population alors que celui des errants (classe 2/2) ne regroupe que 32,5% de la population. Ces derniers sont jeunes, ont occupé des emplois précaires, manquent de gîte et déménagent souvent.

VARIABLES	MODALITES CARACTERISTIQUES
CLASSE 1 / 2	
NOMBRE DE DEMENAGEMENT(S)	PAS ET 1 FOIS
NOMBRE DE JOUR(S) SANS GITE	reponse manquante
MANQUE DE GITE	NON
CONSOMMATION DE DROGUES	NON
AGE	DE 50 A 59
DUREE DU DEUXIEME EMPLOI	98 MOIS ET PLUS
SCOLARITE	QUATRE A 7 ANS
AGE	DE 40 A 49
DUREE DU PREMIER EMPLOI	98 MOIS ET PLUS
DUREE DU DEUXIEME EMPLOI	reponse manquante
RESIDENCE	CHAMBRE
DUREE DU TROISIEME EMPLOI	98 MOIS ET PLUS
AGE	DE 60 ET PLUS
DERNIER JOB STEADY	PLUS DE 15 ANS
DERNIER JOB STEADY	DE 6 A 10 ANS
DEUXIEME EMPLOI	reponse manquante
DUREE DU PREMIER EMPLOI	DE 61 A 97 MOIS
DUREE DU TROISIEME EMPLOI	reponse manquante
PRINCIPAL PROBLEME	SANTE
TROISIEME EMPLOI	reponse manquante
DUREE DU PREMIER EMPLOI	DE 37 A 60 MOIS
PREMIERE SOURCE DE REVENU	BIEN-ETRE SOCIAL
CLASSE 2 / 2	
MANQUE DE GITE	OUI
AGE	DE 0 A 29
NOMBRE DE DEMENAGEMENT(S)	4 FOIS ET PLUS
DUREE DU TROISIEME EMPLOI	DE ZERO A 6 MOIS
DUREE DU PREMIER EMPLOI	DE ZERO A 6 MOIS
CONSOMMATION DE DROGUES	DE TEMPS EN TEMPS
NOMBRE DE DEMENAGEMENT(S)	2 ET 3 FOIS
DUREE DU DEUXIEME EMPLOI	DE ZERO A 6 MOIS
NOMBRE DE JOUR(S) SANS GITE	DE 6 A 10
SCOLARITE	DIX A DOUZE ANS
RESIDENCE	REFUGE
NOMBRE DE JOUR(S) SANS GITE	DE 1 A 5
AGE	DE 30 A 39
NOMBRE DE JOUR(S) SANS GITE	DE 21 A 35
NOMBRE DE JOUR(S) SANS GITE	PLUS DE 3 MOIS
PREMIERE SOURCE DE REVENU	PAS DE REVENU
DERNIER JOB STEADY	MOINS DE 2 ANS
NOMBRE DE JOUR(S) SANS GITE	DE 11 A 20
CONSOMMATION DE DROGUES	SOUVENT
DEUXIEME EMPLOI	AUTRE EMPLOI
RESIDENCE	PENSION
RESIDENCE	reponse manquante

VARIABLES	MODALITES CARACTERISTIQUES
CLASSE 2 / 3	
DUREE DU DEUXIEME EMPLOI	reponse manquante
DEUXIEME EMPLOI	reponse manquante
DUREE DU TROISIEME EMPLOI	reponse manquante
TROISIEME EMPLOI	reponse manquante
DERNIER JOB STEADY	JAMAIS EU
DUREE DU PREMIER EMPLOI	reponse manquante
SCOLARITE	ZERO A TROIS ANS
DUREE DU PREMIER EMPLOI	98 MOIS ET PLUS
PREMIER EMPLOI	reponse manquante
NOMBRE DE DEMENAGEMENT(S)	PAS ET 1 FOIS
MANQUE DE GITE	reponse manquante
CLASSE 3 / 4	
DEUXIEME EMPLOI	AUTRE EMPLOI
PREMIER EMPLOI	AUTRE EMPLOI
TROISIEME EMPLOI	AUTRE EMPLOI
SCOLARITE	TREIZE ANS ET PLUS
PREMIERE SOURCE DE REVENU	PAS DE REVENU
DERNIER JOB STEADY	DE 2 A 5 ANS
CLASSE 2 / 5	
PREMIER EMPLOI	reponse manquante
DUREE DU PREMIER EMPLOI	reponse manquante
DERNIER JOB STEADY	JAMAIS EU
DEUXIEME EMPLOI	reponse manquante
DUREE DU DEUXIEME EMPLOI	reponse manquante
SCOLARITE	ZERO A TROIS ANS

La classification en trois classes montre clairement que la classe 1/2 se scinde à son tour en deux donnant d'une part la classe 1/3 qui conserve les caractéristiques générales des précaires et, d'autre part, la classe 2/3 qui se particularise en regroupant les précaires ayant occupé soit un très long emploi, soit jamais d'emploi. Ce dernier regroupement qui à première vue peut paraître curieux s'explique par le fait qu'il s'agit d'individus qui prennent tous la modalité valeur manquante pour les variables concernant les deuxièmes et troisièmes emplois. D'ailleurs celui-ci éclate au moment de la coupure en 5 classes qui, pour sa part, conserve les trois classes décrites précédemment et laisse émerger deux petits groupes. L'un de ces groupes (2/5) est issu des précaires ayant occupé un très long emploi ou jamais d'emploi, il s'agit des clochards décrochés, les plus hypothéqués. Ils n'ont jamais eu d'emploi stable et sont sous-scolarisés ou pas scolarisés du tout. L'autre petit groupe (4/5), issu de la classification en 4 classes, se conserve dans la partition en 5 classes et est issu des errants. Il correspond à l'image du clochard classique, poète, marginalisé par choix,

anarchiste, artiste... Ce sont des individus à scolarité élevée, ayant exercés des emplois d'artiste ou des emplois marginaux (modalité: autres emplois), mais des emplois stables.

Ce qui est remarquable dans ces résultats, c'est qu'ils font apparaître que l'image caricaturale qu'on associe spontanément à la réalité clocharde, c'est-à-dire l'image de vieillards errants, un peu fous, mendiant au coin des rues, dormant dans les parcs ne correspond dans la réalité qu'à une minorité des itinérants montréalais. Ceux-ci composent un groupe hétérogène où se retrouvent des hommes présentant des besoins diversifiés qui appellent des services diversifiés, ce dont il faut bien se rendre compte.

Conclusion

Dans le cadre du présent article, on a vu que l'utilisation des méthodes multidimensionnelles a permis de préciser une intuition.

La complémentarité entre méthodes univariées et méthodes multivariées ne doit toutefois pas être envisagée comme devant nécessairement aboutir à la validation des premiers résultats ou d'hypothèses issues de ces premiers résultats. Il pourrait tout aussi bien arriver qu'au contraire les intuitions de départ soient infirmées suite à la réalisation d'analyses multivariées; l'effet combiné d'un ensemble de variables neutralisant la corrélation préalablement observée entre les variables prises deux à deux. Ceci ne remet toutefois pas en cause la validité des techniques. Chacune permet de se faire une idée d'un aspect de la réalité étudiée. Les unes sont complémentaires des autres, et c'est au chercheur qu'il appartient de ne négliger aucune ouverture possible et en évitant les à priori afin de cerner du mieux possible la réalité et de la modéliser quand cela s'avère possible.

Bibliographie

BENZECRI J.P. et al., L'analyse des données, Tomes 1 et 2, DUNOD, Paris, 1973.

BOUROCHE J.M., SAPORTA G., L'analyse des données, Que sais-je ? n° 1854, PUF, 1983.

DIDAY E., LEMAIRE J., POUGET J., TESTU F., Eléments d'Analyse de données, DUNOD, Paris, 1982.

LEBART L., MORINEAU A., TABARD N., Technique de la description statistique. DUNOD, Paris, 1977.

ROY S., Seuls dans la rue, Editions St-Martin, Montréal, 1988.

ر

CONSTRUCTION D'UNE HIERARCHIE CONSENSUS A L'AIDE D'UNE ULTRAMETRIQUE CENTRALE

Guy Cucumel[1]

Département de sociologie, UQAM,
CP 8888, Succ. A, Montréal (Québec)
H3C 3P8

Introduction

Le problème de la comparaison de classifications hiérarchiques se présente lorsque l'on travaille sur plusieurs ensembles de variables, lorsque les classifications sont obtenues à partir de critères d'agrégation différents ou lorsque l'on dispose de variables évoluant dans le temps... Le problème que l'on se pose est le suivant: comment obtenir à partir d'une famille de k hiérarchies définies sur un même ensemble d'objets, une hiérarchie unique représentative, en un sens à préciser, de cette famille. Il s'agit "d'étendre" la notion de moyenne à des structures aussi complexes que les hiérarchies. Le sens de la représentativité de la hiérarchie consensus a souvent été discuté au cours de ces dernières années et a donné lieu à la production d'algorithmes nombreux et très différents, chaque auteur ayant sa propre idée de la représentativité. En fait, celle-ci est souvent liée au type de données et de hiérarchies, indicées ou non, dont on cherche le consensus. Il nous semble illusoire de croire qu'il existe un algorithme général adapté à toutes les situations.

La paternité du premier algorithme de recherche de consensus peut être attribué à Adams (1972) qui adapte son algorithme de recherche de consensus d'arbres phylogénétiques au cas des hiérarchies. Neumann (1983), Stinebrickner (1984), Diday et Cucumel (1988) proposent, pour leurs parts, divers algorithmes s'appliquant sur des hiérarchies indicées. Tous ces algorithmes présentent un inconvénient en ne tenant pas compte des échelles d'indiçage, ce qui, dans certains cas, peut mener à des abhérations; une des k hiérarchies devenant le consensus de façon dictatoriale car les échelles d'indiçage ne sont pas comparables. D'autre part, on constate que certaines méthodes ne tiennent pas compte du poids de chaque hiérarchie, c'est-à-dire que si une même hiérarchie est présente plusieurs fois dans le k-uplet, elle ne sera prise en compte qu'une seule fois. Or, dans la pratique, ce cas risque de se produire assez fréquemment dans la mesure où il nous semble raisonnable de rechercher un consensus sur des hiérarchies pas trop "éloignées", afin que celui-ci ait un sens.

Nous proposons une autre approche, basée sur la recherche d'une hiérarchie centrale à l'aide d'une ultramétrique centrale. Cette approche est une méthode de résolution plus qu'un algorithme et débouche sur la constitution d'une famille de hiérarchies de topologies identiques, mais d'indiçages différents.

1. Définitions et notations

Le présent paragraphe a pour but de fixer les notations et de rappeler les définitions essentielles.

---

[1] Chercheur au département de Sociologie de l'UQAM, membre du GRAPPP (Groupe de Recherche et d'Analyse sur les Politiques et les Pratiques Pénales).

On note $\Omega=\{w_1,w_2,...,w_n\}$ un ensemble fini de n objets, $\mathcal{P}(\Omega)$ l'ensemble des parties de Ω, $(H_1,H_2,...,H_k)$ un k-uplet de hiérarchies définies sur Ω et $\mathcal{H}$ l'ensemble des hiérarchies définies sur Ω.

On appelle **hiérarchie H** tout ensemble de parties de Ω (appelées paliers, nœuds ou classes) tel que :

i) $\Omega \in H$

ii) $\forall\, i \in \{1, 2, ..., n\}, \{w_i\} \in H$

iii) $\forall\, (h_1, h_2) \in H \times H, h_1 \cap h_2 \in \{\emptyset, h_1, h_2\}$

On peut associer à chaque palier une valeur numérique qui correspond sur la représentation graphique à la hauteur du palier et permet d'avoir une idée du niveau d'agrégation des deux parties de la hiérarchie qu'il réunit. Ceci nous amène à rappeler la notion de hiérarchie indicée.

On appelle **hiérarchie indicée (H, f)** tout couple tel que :

i) H est une hiérarchie

ii) f est une application de $\mathcal{P}(\Omega)$ dans $\mathbb{R}^+$ vérifiant

a) $\forall\, i \in \{1, 2, ..., n\}, f(\{w_i\}) = 0$

b) $\forall\, (h_1, h_2) \in H \times H, h_1 \subseteq h_2 \Rightarrow f(h_1) \leq f(h_2)$

Enfin, on utilisera dans cette recherche les deux notions suivantes:

On appelle **indice de dissimilarité d** une application de $\Omega \times \Omega \rightarrow \mathbb{R}^+$ telle que :

i) $\forall\, (w_i, w_{i'}) \in \Omega \times \Omega\,,\ d(w_i, w_{i'}) = d(w_{i'}, w_i)$

ii) $\forall\, w_i \in \Omega,\ d(w_i, w_i) = 0$

Si de plus d vérifie les deux propriétés suivantes:

i) $\forall\, (w_i, w_{i'}) \in \Omega \times \Omega\,,\ d(w_i, w_{i'}) = 0 \Leftrightarrow w_i = w_{i'}$

ii) $\forall\, (w_i, w_{i'}, w_{i''}) \in \Omega \times \Omega \times \Omega,\ d(w_i, w_{i''}) \leq \max\,(d(w_i, w_{i'}), d(w_{i'}, w_{i''}))$

alors d est une **ultramétrique**.

Les relations et les propriétés de ces différents indices sont assez nombreuses, parmi celles-ci on note le résultat remarquable suivant: étant donnée une hiérarchie indicée (H, f), il est toujours possible d'induire à partir de celle-ci une ultramétrique σ pour laquelle la dissimilarité entre deux objets est égale à la hauteur du plus bas palier qui les contient. Formellement, on a:

$$\sigma(w_i, w_{i'}) = \min_{h \in H} \{f(h),\ w_i \in h \text{ et } w_{i'} \in h\}$$

2. Distances entre hiérarchies

Boorman et Olivier (1973) ont défini des mesures générales de distance entre deux hiérarchies basées sur les différences entre paires de partitions, chacune issue d'une des deux hiérarchies.

On propose ici de mesurer la distance entre hiérarchies par la distance entre les ultramétriques associées. Par le théorème de Benzecri-Johnson (1967), on sait qu'il existe une bijection entre l'ensemble des hiérarchies indicées définies sur un ensemble Ω et l'ensemble des ultramétriques définies sur Ω. Autrement dit, il est absolument équivalent de se donner une ultramétrique ou une hiérarchie indicée. L'ultramétrique correspond à la dissimilarité induite par la hiérarchie indicée.

Le problème se repose donc en termes de distances entre ultramétriques. Plusieurs auteurs dont Sokal et Rohlf (1962), Hartigan (1967), Gower (1971) et Hubert (1974) ont proposé des mesures de distorsion entre une ultramétrique et une dissimilarité. Dans notre approche, considérant qu'une ultramétrique est une dissimilarité particulière, on utilisera la sommes des carrés développée par Hartigan.

La distance entre deux hiérarchies indicées (H, f) et (H', f') est donnée par l'expression suivante:

$$D[(H, f), (H', f')] = \sum_{(w_i, w_{i'}) \in \Omega \times \Omega} [u(w_i, w_{i'}) - u'(w_i, w_{i'})]^2$$

où u et u' sont les deux ultramétriques induites par (H, f) et (H', f').

3. Position du problème

On décide d'aborder le problème dans les termes suivants: considérant une famille de hiérarchies (H_j, f_j) pour $j = 1, ..., k$, on recherche la hiérarchie indicée (H^*, f^*) à distance minimum de chacune des hiérarchies indicées (H_j, f_j). En d'autres termes, il s'agit de trouver la hiérarchie centrale du k-uplet de hiérarchies. On considérera la hiérarchie centrale comme représentative du k-uplet.

On appelle **hiérarchie indicée centrale** la hiérachie (H^*, f^*) vérifiant :

$$\sum_{j=1}^{k} D[(H^*, f^*), (H_j, f_j)]) = \min_{(H, f)} (\sum_{j=1}^{k} D[(H, f), (H_j, f_j)])$$

où D est la distance entre hiérarchies définie au paragraphe 2.

Le problème peut alors se poser de la façon suivante :

On cherche $u^* \in \mathcal{U}$ (ensemble des ultramétriques définies sur Ω) réalisant le minimum sur $\mathcal{U}$ de:

$$\sum_{j=1}^{k} (\sum_{(w_i, w_{i'}) \in \Omega \times \Omega} [u(w_i, w_{i'}) - u_j(w_i, w_{i'})]^2)$$

où u appartient à $\mathcal{U}$. u^* apparaît naturellement comme une **ultramétrique centrale** au sens des moindres carrés.

Afin d'alléger les notations, $(w_i, w_{i'}) \in \Omega \times \Omega$ sera noté $(w_i, w_{i'})$.

La quantité à minimiser devient donc :

$$\sum_{j=1}^{k} \left(\sum_{(w_i, w_{i'})} [u(w_i, w_{i'}) - u_j(w_i, w_{i'})]^2 \right)$$

soit encore: $$\sum_{(w_i, w_{i'})} \left(\sum_{j=1}^{k} [u(w_i, w_{i'}) - u_j(w_i, w_{i'})]^2 \right) \qquad (1)$$

Suivant cette approche, toutes les hiérarchies ont le même poids, à condition que les échelles d'indiçage soient comparables.

Afin de s'affranchir des problèmes d'échelles, on travaille sur des **hiérarchies normées**, c'est-à-dire des hiérarchies dont l'indiçage est compris entre 0 et 1.

Remarquons que toute hiérarchie indicée peut être normée, il suffit de diviser l'indiçage de chacun des paliers par f (Ω).

Sans perte de généralité, on peut considérer que la résolution du problème s'effectue sur un k-uplet de hiérarchies normées, i.e. que les ultramétriques u_j pour j=1,..., k prennent leurs valeurs sur l'intervalle [0, 1].

4. La méthode

Le problème de minimisation va nécessairement trouver sa solution sur l'intervalle [0,1[, ce qui pose la question de l'interprétation de l'indiçage obtenu. Afin de contourner cette difficulté, on propose d'effectuer un deuxième changement d'échelle afin de se ramener à une échelle interprétable dans les mêmes termes que l'une des échelles initiales. Autrement dit, on obtient k hiérarchies de même topologie, mais d'indiçages différents. On peut ainsi comparer visuellement et interpréter la hiérarchie consensus par rapport à chacune des hiérarchies du k-uplet.

La méthode que l'on propose pour résoudre le problème est alors la suivante :

-résoudre le problème (1) sur des hiérarchies normées et construire la hiérarchie indicée correspondant à l'ultramétrique trouvée;

-se ramener à l'une des échelles d'indiçage initiales pour interpréter la solution trouvée. Pour cela, on effectue de nouveau un changement d'échelle de manière à ce que le palier le plus haut de la hiérarchie centrale ait le même indiçage que le palier le plus haut de la hiéarchie indicée par l'échelle initiale.

5. Solution du problème

Le problème (1) se transforme aisément en minimiser

$$k \sum_{(w_i, w_{i'})} [\, u(w_i, w_{i'}) - \frac{1}{k} \sum_{j=1}^{k} u_j(w_i, w_{i'}) \,]^2 - \frac{1}{k^2} \left(\sum_{j=1}^{k} u_j(w_i, w_{i'}) \right)^2 + \frac{1}{k} \left(\sum_{j=1}^{k} u_j^2(w_i, w_{i'}) \right)$$

comme $\frac{1}{k^2} \left(\sum_{j=1}^{k} u_j(w_i, w_{i'}) \right)^2 + \frac{1}{k} \left(\sum_{j=1}^{k} u_j^2 (w_i, w_{i'}) \right)$ est constant, la quantité à minimiser devient :

$$\sum_{(w_i, w_{i'})} [\, u(w_i, w_{i'}) - \frac{1}{k} \sum_{j=1}^{k} u_j(w_i, w_{i'}) \,]^2$$

Posons $d(w_i, w_{i'}) = \frac{1}{k} \sum_{j=1}^{k} u_j(w_i, w_{i'})$

On doit donc résoudre le problème :

$$\min_{u \in \mathcal{U}} \sum_{(w_i, w_{i'})} [\, u(w_i, w_{i'}) - d(w_i, w_{i'}) \,]^2 \qquad (2)$$

Il est facile de démontrer que d définie par :
$\forall\, (w_i, w_{i'}) \in \Omega \times \Omega,\ d(w_i, w_{i'}) = \frac{1}{k} \sum_{j=1}^{k} u_j(w_i, w_{i'})$
est une dissimilarité sur Ω.

Le problème que l'on à résoudre est donc de trouver l'ultramétrique la plus proche d'une dissimilarité donnée au sens des moindres carrés.

Le problème (2) peut encore s'écrire sous la forme d'un problème d'optimisation sous contraintes. En effet, le fait que l'on cherche à une minimisation sur l'ensemble des ultra-métriques permet de réécrire le problème (2) sous la forme suivante :

$$\begin{cases} \min \sum_{(w_i, w_{i'})} [\, u\,(w_i, w_{i'}) - d(w_i, w_{i'}) \,]^2 \\ \text{sous les contraintes} \\ \forall\, (w_i, w_{i'}, w_{i''}) \in \Omega \times \Omega \times \Omega,\ u\,(w_i, w_{i''}) \leq \max\,(u\,(w_i, w_{i'}), u\,(w_{i'}, w_{i''})) \end{cases}$$

Ce problème de minimisation sous contraintes ne peut être résolu comme tel. Par contre, il peut être décomposé en plusieurs sous problèmes en réécrivant les contraintes de façon à les rendre

linéaires tout en respectant les inégalités imposées par le fait que u doit être une ultramétrique. Chaque sous problème devient alors un problème classique de minimisation sous contraintes.

La solution du problème (2) est alors donnée par le sous problème minimisant

$$\sum_{(w_i,\, w_{i'})} [\, u(w_i, w_{i'}) - d(w_i, w_{i'}) \,]^2 \ .$$

Si ceci prouve qu'il existe une solution, par contre il est clair qu'elle ne peut-être calculée de cette façon. En effet, il existe autant de sous problèmes que de hiérarchies sur n éléments. Ce nombre est égal à $n!\ (n-1)/\ 2^{n-1}$ qui correspond au nombre de hiérarchies binaires (Frank et Svensson, 1981).

Sokal et Rohlf (1962), puis Sokal et Sneath (1963) furent les premiers à poser ce problème. Devant la complexité combinatoire, plusieurs algorithmes de recherche d'un optimum local ont été proposés. Une première famille d'algorithmes repose sur des opérations locales consistant à échanger ou à supprimer certains sommets de la représentation arborescente des dissimilarités ultramétriques (c.f. Hartigan, 1967 ; Caroll et Chang, 1973). Une deuxième famille d'algorithmes utilise une méthode de gradient (c.f. Caroll et Pruzansky ,1975).

On peut signaler également les études empiriques de Sokal et Rholf (1962) , Boyce (1969), Jambu (1972) et Farris (1973) qui montrent que parmi la famille des algorithmes ascendants hiérarchiques de Lance et Williams (1967), l'algorithme de la moyenne est celui qui conduit le plus fréquemment à la plus petite valeur du critère à minimiser. On trouve dans la thèse de Davy de Virville (1972) et dans l'article de Florens et Lochard (1976) une étude des propriétés mathématiques d'un optimum global.

Une solution du problème est donnée par Chandon, Lemaire et Pouget (1980). Sans détailler l'algorithme, rappelons simplement, qu'il consiste à remplacer les n(n-1)/2 valeurs de la dissimilarité d par n-1 valeurs de l'ultramétrique u, en construisant simultanément n-1 classes d'équivalence emboitées, c'est-à-dire une hiérarchie de partitions, et en déterminant les n-1 valeurs de l'ultramétrique associée. Cet algorithme ascendant de branchement généralise celui de la moyenne, ce qui explique qu'il fournisse, une remarquable solution approchée. La différence est qu'à chaque noeud, on ne se limite pas à la fusion des deux classes h et h' minimisant l'indice d'agrégation du lien moyen $\delta(h, h')$, on explore tous les regroupements tombant dans un intervalle d'admissibilité, dont la borne inférieure est la moyenne des dissimilarités entre les paires d'objets déjà agrégés et la borne supérieure la moyenne entre les paires d'objets non encore agrégés.

6. Exemple

Considérons les trois hiérarchies H_1, H_2 et H_3 de la figure 1. La colonne de droite présente les trois hiérarchies consensus, de même topologie mais d'indiçages différents, issues de la méthode expliquée au paragraphe 4.

7. Conclusion

L'algorithme de Chandon, Lemaire et Pouget (1980) permet de trouver la solution optimale du problème. Cependant, comme tout algorithme de branchement, il implique l'exploration de nombreuses branches, et n'est efficace que si le nombre d'objets de Ω est inférieur à 15. Chandon, Lemaire et Pouget propose un autre algorithme divisif rapide, permettant d'arriver à une solution

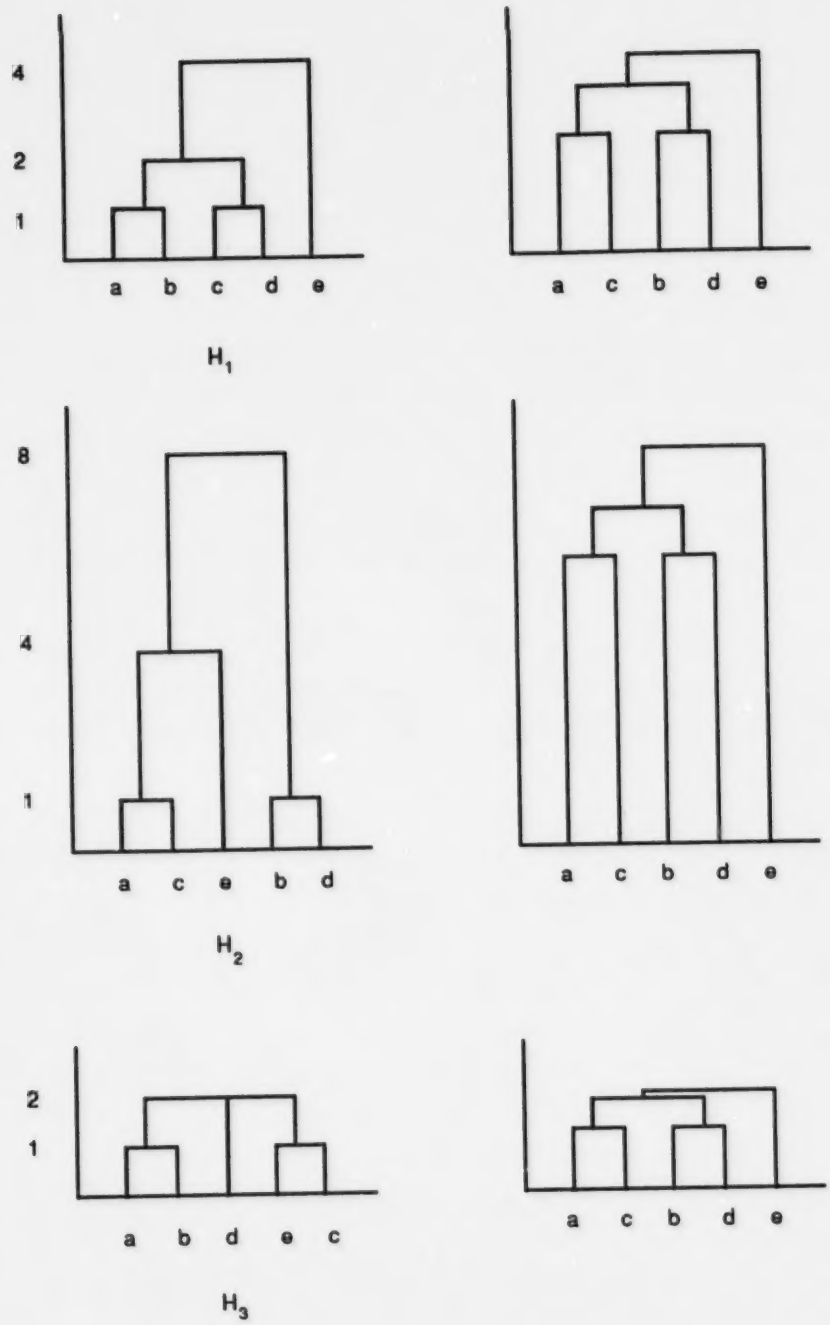


figure 1

non nécessairement optimale, mais satisfaisante bien que présentant parfois des inversions dans la hiérarchie construite. Dans la pratique l'algorithme du lien moyen conduit à une bonne solution approchée.

L'intérêt de la méthode que nous avons développée ici est d'être indépendante du type d'algorithme ou d'heuristique utilisé pour calculer l'ultramétrique la plus proche - au sens des moindres carrés - de la dissimilarité d définie au paragraphe 5. Plus généralement elle est indépendante de la distance D choisie pour mesurer la distance entre hiérarchies, pour autant que celle-ci soit exprimée en termes d'écart entre une dissimilarité et une ultramétrique. La prolongation de la présente recherche passe par l'étude d'autres distances de ce type.

Bibliographie

ADAMS, E.N.,III (1972) : "Consensus Techniques and the Comparison of Taxonomic Trees", *Systematic Zoology*, 21, 390-397.

BOORMAN , S.A., OLIVER, D.C. (1973) : Metrics on Spaces of Finite Trees, *J. Math. Psychol.*, 10, 26-59.

BOYCE, A.J. , (1969) : Mapping Diversity : A Comparative Study of Some Numerical Methods, in *Numeriacl Taxonomy*, COLE ed., Academic Press.

CARROLL, J.D., CHANG, J.J., (1973) : A Method for Fitting a Class of Hierarchical tree Structure Models to Dissimilarities Data and Its Application to Some "Body Parts" Data of Miller's, *Proceedings*, 81st Annual Convention A.P.A.

CARROLL, J.D., PRUZANSKY, S., (1975) : Fitting of Hierarchical Tree Structure (HTS) Models, Mixtures of HTS, and Hybrid Models, via Mathematical Programming and Alternating Least Squares, in *Theory, Methods and Applications of Multidimensional Scaling and Related Techniques*, U.S. Japan Seminar, University of California.

CHANDON, J.L., DE SOETE, G., (1984), Fitting a least squares ultrametric to dissimilarity data : approximation versus optimization, in *Data Analysis and Informatics III*, E.Diday et al. (ed.), North-Holland, Amsterdam.

CHANDON, J.L., LEMAIRE J., POUGET J., (1980) : Construction de l'ultramétrique la plus proche d'une dissimilarité au sens des moindres carrés, *R.A.I.R.O. Recherche Opérationnelle*, 14, 157-170.

DAVY de VIRVILLE, M., (1972) : *L'analyse des proximités*, Thèse de 3ème cycle, Université de Paris VI.

DIDAY, E., CUCUMEL G., (1988) : Compatibility and Consensus in Numerical Taxonomy, in *Proceedings of the 9th international conference on pattern recognition*, IEEE, Washington.

FARRIS, J.S., (1973) : *On comparing the shape of taxonomic trees*, Syst. Zool. 22, 50-54.

FLORENS, D., LOCHARD, E.O., (1976)*Taxonomie automatique : la sous-dominante dans son cadre général et la "plus proche du sens des moindres carrés"*, Cahierdu Bureau Universitaire de Recherche opérationneile.

GOWER, J.C., (1971) : Statistical Methods of Comparing Different multivariate Analyses of the Same Data Set, in *Mathematics in the Archaelogical and Historical Sciences* (F.R. Hodson, D.G. Kendall and P. Tautu, eds.), 138-149, Edinburgh: Edinburgh University Press.

HARTIGAN, J.A., (1967) : Representation of Similarity Matrices by Trees, *J. Amer. Stat. Assoc.*, vol. 62, 1140-1158.

HUBERT, L., Approximate Evaluation Techniques for the Single-link and Complete-link Hierarchical Clustering Procedures, *J. Amer. Stat. Assoc.*, vol. 69, 698-704.

JAMBU, M., (1972) :*Techniques de classification automatique appliquée à des données de Sciences humaines*, Thèse de Doctorat de 3ème cycle, Paris.

JARDINE, N., SIBSON, R., (1971) : *Mathematical Taxonomy*, London, Wiley.

JOHNSON, S.C., (1967) : Hierarchical clustering Schemes, *Psychométrica*, vol. 32, 241-254.

LANCE, G.N., WILLIAMS, W.T., (1967) : A General Theory of Classification Sorting Strategies: 1. Hierarchical Systems; 2. Clusterig Systems, *Computer Journal*, vol. 9-10, 373-380.

NEUMANN, D.A., (1983) : "Faithful Consensus Methods for n-Trees", *Math. Biosci.*, 63, 271-287.

SOKAL, R.R., ROHLF, F.J., (1962) : The Comparison of Dendrograms by Objective Methods, *Taxon*, Taxon, vol.11, 33-40.

SOKAL, R.R., SNEATH, P.H.A., (1963) : *Principles of Numerical Taxonomy*, Freeman.

STINEBRICKNER, R., (1984) : "s-Consensus Trees and Indices", *Bull. of Math. Biology*, 46, 923-935.

STINEBRICKNER, R., (1984) : "An Extension of Intersection Methods from Trees to Dendrograms", *Syst. Zool.*, 33, 381-386.

Modèles linéaires généralisés: analyse des résidus pour des données provenant d'une distribution binomiale

par Mario Lafond

Résumé

Les modèles linéaires généralisés sont un outil statistique permettant de modéliser des observations provenant de distributions comprises dans la famille exponentielle (par exemple, la loi normale, la loi gamma, la loi de Poisson et la loi binomiale).

Le modèle binomial, tout comme la régression classique, mérite une attention particulière pour fin de validation. De par la structure innée de ce modèle, plusieurs définitions de résidus sont possibles. Ces différentes définitions seront présentées et discutées. De plus, on verra comment les résidus peuvent nous permettre de valider plusieurs des composantes du modèle.

1. Rappel sur les modèles linéaires généralisés

Pour construire un modèle linéaire généralisé, on considère un ensemble de paramètres $\beta_1, \ldots, \beta_r$ et un ensemble de variables explicatives X_j tels que

$$g(\mu_i) = (X\beta)_i = \sum_j x_{ij}\beta_j$$

où la fonction $g(.)$ est appelée *fonction de lien.*

Ainsi on remarque les trois éléments essentiels des modèles linéaires généralisés,

a. la *composante aléatoire* constituée du vecteur Y dont les éléments Y_i sont des variables aléatoires de densité f_{Y_i} (par exemple, la loi binomiale);

b. la *composante systématique*

$$\eta = X\beta;$$

c. le *lien* entre ces deux composantes

$$\eta = g(\mu).$$

Les estimateurs de ces modèles ne peuvent pas être donnés de façon explicite; c'est pourquoi ils sont résolus par une méthode itérative: la méthode de Newton. Grâce à cette méthode, on trouve que l'estimateur de β, à savoir $\hat{\beta}$, est donné par

$$\hat{\beta} = (X^t W X)^{-1} X^t W Z$$

où W est une matrice diagonale de pondération dont les éléments non nuls sont définis comme étant

$$w_{ii} = \frac{1}{V(\mu_i)(g'(\mu_i))^2},$$

où $V(.)$ représente la fonction de variance associée à la distribution et Z représente la variable réponse transformée (développement en série de Taylor) comme suit

$$Z = g(\mu_i) + (y_i - \mu_i)g'(\mu_i).$$

2. Notations spécifiques à la distribution binomiale

La fonction de densité de la distribution binomiale est donnée par

$$f_Y(y;n,p) = \exp\{\ln\binom{n}{p} + n\ln(1-p) + y\ln\frac{p}{1-p}\}, \quad ,y = 0,\ldots,n.$$

Pour sa part, la fonction de variance associée à cette distribution est habituellement donnée par

$$V(\mu_i) = \frac{\mu_i(1-\mu_i)}{n_i}, \quad \mu_i \in (0,1).$$

Quant à l'estimation de l'écart-type, il est donné par

$$s_i = \sqrt{\frac{\hat{p}i(1-\hat{p}_i)}{n_i}}.$$

3. Calcul des différentes catégories de résidus

3.1 *Résidu linéaire*

Le résidu linéaire représente, comme son nom le laisse présager, une relation linéaire entre la valeur observée et la valeur espérée, c'est-à-dire

$$R_L(y,\theta) = \frac{y - E_\theta[Y]}{S_\theta(Y)}.$$

Parmi cette vaste catégorie de résidus, on retrouve un des résidus les plus répandus: le résidu de Pearson. Ce dernier se définit comme étant

$$R_P(i) = \frac{y_i - \hat{y}_i}{s_i\sqrt{(1-h_{ii})}}$$

où h_{ii} est l'élément (i,i) de la matrice de projection associée aux modèles linéaires généralisés, c'est-à-dire

$$H = W^{\frac{1}{2}} X (X^t W X)^{-1} X^t W^{\frac{1}{2}}.$$

Finalement, les propriétés du résidu linéaire, en particulier celui de Pearson, sont

. souligne l'écart entre la valeur observée et la valeur prédite;

. permet le calcul du khi-carré, $\chi^2 = \sum_i R_P(\cdot)$;

. facile de calcul;

. interprétation facile.

Dans le cas où l'on traite des individus provenant d'une distribution binomiale, le résidu de Pearson nous est donné par

$$R_P(i) = \frac{n_i^{\frac{1}{2}}(p_i - \hat{p}_i)}{\sqrt{\hat{p}_i(1-\hat{p}_i)(1-h_{ii})}}.$$

3.2 *Résidu associé à une transformation*

Par suite d'une transformation de la variable dépendante Y_i, on obtient la forme générale du résidu associé à une transformation,

$$R_T(y;\theta) = \frac{t(y) - E_\theta[t(Y)]}{S_\theta(t(Y))}.$$

Parmi cette catégorie de résidus, on retrouve plusieurs définitions de résidus ayant des buts identiques: associer un résidu spécifique à une distribution spécifique. Parmi

les résidus répondant à ce critère, on note premièrement le résidu d'Anscombe qui est défini à partir de la fonction de variance. Ainsi on doit d'abord calculer

$$A(\mu) = \int \frac{d\mu}{V^{\frac{1}{3}}(\mu)}$$

qui sert de point de départ au calcul du résidu d'Anscombe qui se définit comme

$$R_A(i) = \frac{A(y_i) - A(\hat{y}_i)}{S_i A'(y_i)}.$$

De plus, il peut être judicieux d'ajuster ce résidu par rapport au troisième moment centré. Ainsi le résidu d'Anscombe devient

$$R_{AA} = R_A + \frac{p_3(\hat{\theta})}{6}.$$

Particulièrement en ce qui concerne les observations provenant d'une distribution binomiale, le résidu d'Anscombe est donné par

$$R_A = \frac{\sqrt{n_i}(A(p_i) - A(\hat{p}_i))}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$$

où la quantité $A(p_i) - A(\hat{p}_i)$ correspond à

$$\int_{\hat{p}_i}^{p_i} s^{-\frac{1}{3}}(1-s)^{-\frac{1}{3}}\, ds.$$

De plus, le troisième moment centré est donné, pour la loi binomiale, par

$$p_3(\hat{\theta}) = \frac{(1-2\hat{p}_i)\sqrt{\hat{p}_i(1-\hat{p}_i)}}{n_i}.$$

En plus du résidu défini par Anscombe, il existe d'autres résidus associés à une transformation. Pour le cas binomial, il est possible de définir un résidu stabilisant la

variance. Ce résidu, populaire auprès des utilisateurs de modèles binomials est donné par

$$R_{SV} = 2\sqrt{n_i}(arcsin\sqrt{p_i} - arcsin\sqrt{\hat{p}_i}).$$

Pour tous les résidus associés à des transformations y compris celui d'Anscombe et celui stabilisant la variance, on retrouve les propriétés suivantes:

. respect de la loi;

. calcul ardu pour certaine distribution;

. n'existe pas pour toutes les distributions.

3.3 *Résidu de la déviance*

La déviance représente la statistique permettant de vérifier l'ajustement du modèle proposé. La déviance échantillonale est définie comme étant la différence entre le logarithme de la fonction de vraisemblance de modèle sélectionné et celui du modèle où l'ajustement serait parfait

$$Dev = -2\sum_{i=1}^{n}(L(\hat{y}_i; y_i) - L(y_i; y_i)) = \sum_{i=1}^{n} Dev_i$$

où $L(.;.)$ représente le logarithme de la fonction de vraisemblance.

A partir de cette définition de la déviance échantillonale, on peut définir le résidu de la déviance comme étant

$$R_D(y;\theta) = signe(y - \hat{y})\sqrt{2L(\hat{y}; y) - 2L(y; y)}.$$

De manière moins abstraite, on trouve que

$$R_D(i) = signe(y_i - \hat{y}_i)\sqrt{Dev_i}.$$

Pour le modèle binomial, le résidu de la déviance se ramène à

$$R_D(i) = signe(p_i - \hat{p}_i)\sqrt{2n_i p_i \ln\frac{p_i}{\hat{p}_i} + 2n_i(1-p_i)\ln\frac{(1-p_i)}{(1-\hat{p}_i)}}.$$

Les propriétés du résidu de la déviance sont

. respect de la loi;

. provient de la fonction de vraisemblance;

. facile de calcul.

3.4 *Choix de la meilleure définition de résidus*

A partir des trois catégories de résidus préalablement définis, on en retrouve deux qui existent pour toutes les distributions de la famille exponentielle y compris la distribution binomiale. Ainsi parmi les résidus qui pourraient servir de point de départ pour la construction d'un résidu global, on dénote le résidu de Pearson et le résidu de la déviance. A partir de ces deux résidus, on peut élaborer un résidu conservant le maximum d'information. Ainsi, on peut facilement construire un résidu qui représentera une combinaison linéaire de ces deux résidus. En se servant de h_{ii} comme élément de pondération, on définit le carré du résidu global comme suit

$$R_G^2(i) = (1 - h_{ii})R_D^2(i) + h_{ii}R_P^2(i),$$

ce qui nous permet de définir le résidu global comme étant

$$R_G(i) = signe(y_i - \hat{y}_i)\sqrt{R_G^2(i)}.$$

Ce résidu possède ainsi des éléments majeurs tels que $R_D(i)$, le résidu de la déviance, $R_P(i)$, le résidu de Pearson et h_{ii}, l'élément (i,i) de la matrice de projection.

3.5 *Utilisation des résidus pour fin de validation*

Tous les résidus définis ci-dessus peuvent servir de point de départ pour la validation du modèle. Parmi les éléments que l'on peut valider à l'aide des résidus, on note la fonction de lien, la fonction de variance et l'influence de chacun des individus.

Pour valider la fonction de lien, on a, tout simplement, à vérifier le comportement du graphique des résidus (peut importe la définition retenue) en fonction des valeurs prédites par le modèle. Si ce graphique comporte une allure particulière, alors on peut se questionner sur notre choix de fonction de lien. Grâce à ce graphique, on peut aussi percevoir un problème quant à la forme de la fonction de variance établie pour ce modèle. Il va de soi que la validation de la fonction de lien et de la fonction de variance peut, et doit, aussi être validé par d'autres techniques efficaces. Parmi les techniques de validation de la fonction de lien, on note la présence des familles de fonctions de lien qui permettent de choisir la meilleure fonction de lien possible parmi un grand éventail de fonctions.

Pour ce qui est de la détection des données influentes à l'aide des résidus, il suffit de vérifier si certains des résidus présentent des valeurs anormalement élévées. Il faut cependant avoir une bonne connaissance de son modèle puisqu'un résidu élévé n'indique pas nécessairement une observation influente. Ainsi il est fort probable, surtout pour le modèle binomial, que l'estimation de l'écart-type soit près de zéro ce qui aura tendance à augmenter indûment la valeur du résidu. Tout comme pour les fonctions de lien et de variance, il existe des mesures d'influence plus poussées qui permettront un autre regard sur les observations. Ces mesures d'influence ressemblent à celles conçues pour le modèle classique.

Références utiles

[1] Anscombe, F.J., *Contribution to the discussion of Hotelling's paper*, J. Roy. Statist. Soc. **15** (1953), 229–230.

[2] Anscombe, F.J., Tukey, J.W., *The examination and analysis of residuals*, Technometrics **5** (1963), 141–60.

[3] Cook, R.D., Weisberg, S., "Residuals and Influence in Regression," Chapman and Hall, New York, 1982.

[4] Landwehr, J.M., Pregibon, D., Shoemaker, A.C., *Graphical methods for assessing logistic regression models*, J. Amer. Statist. Assoc. **79** (1984), 61–83.

[5] McCullagh, P., Nelder, J.A., "Generalized Linear Models (second edition)," Chapman and Hall, New York, 1989.

[6] Pierce, D.A., Schafer, D.W., *Residuals and generalized linear models*, J. Amer. Statist. Assoc. **81** (1986), 977-86.

[7] Pregibon, D., *Logistic regression diagnostics*, Ann. Statist. **9** (1981), 705–24.

[8] Williams, D.A., *Generalized linear model diagnostics using deviance and single point deletions*, Appl. Statist. **36** (1987), 181–91.

La Comparaison Statistique de Classifications Hiérarchiques: Evaluation de Trois Tests de Comparaison par Permutations

François-Joseph Lapointe
Département de Sciences biologiques, Université de Montréal
C.P. 6128, Succursale A, Montréal, Québec, H3C 3J7, CANADA

Résumé: Les arbres utilisés en taxonomie numérique sont généralement représentés sous la forme de classifications hiérarchiques ou dendrogrammes. Pour chaque dendrogramme, il existe une matrice cophénétique unique représentant les distances entre chaque paire d'objets. Il est donc possible de comparer des dendrogrammes à l'aide de ces matrices cophénétiques. Trois différentes méthodes de comparaison seront présentées: le test de Mantel (1967), la méthode de Shao et Sokal (1986) et la procédure à doubles permutations de Lapointe et Legendre (1990). Ces trois tests appartiennent à la famille des tests par permutations et permettent de tenir compte d'aspects différents lors des comparaisons de dendrogrammes. Une comparaison théorique et empirique de ces trois méthodes est effectuée. Les résultats suggèrent l'utilisation préférentielle du test de Lapointe et Legendre pour la comparaison de classifications hiérarchiques.

Le concept d'arbre en mathématique

Un *graphe* G (Fig.1) est composé d'un ensemble fini S={s,t,u,v,w,x} de *sommets* reliés par des *branches* ou *arêtes* formant un ensemble fini A={st,su,tu,uv,uw,vx,wx}.

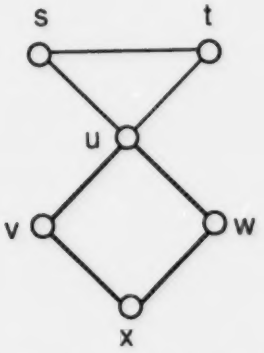


Fig. 1. Représentation d'un graphe.

Un *chemin* c est une suite de branches {su,uw} reliant deux sommets s et w qui sont les *extrémités* de c. Deux sommets t et u sont dit *adjacents* si une arête unique {tu} les relie. Un *cycle* est un chemin {uv,vx,xw,wu} dont les extrémités coïncident. Le graphe G est *connexe* lorsqu'il existe au moins un chemin entre chaque paire de sommets de G. Un *arbre* (Fig. 2) est défini comme un graphe connexe sans cycle, c'est-à-dire, un graphe ou toutes les

paires de sommets sont reliées par un chemin et un seul (Barthélemy et Guénoche 1988).

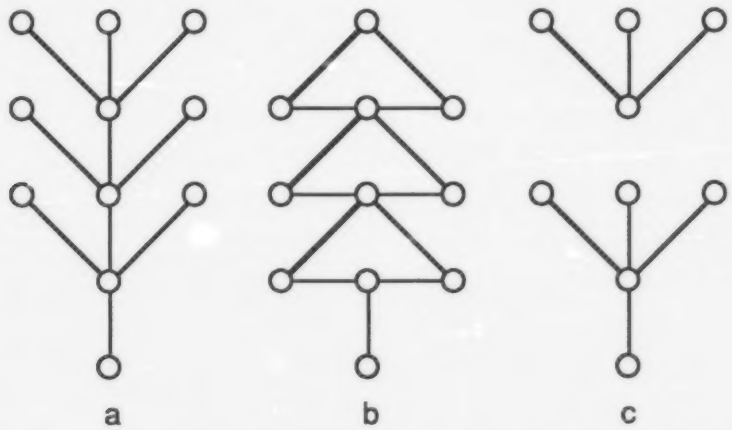


Fig.2. Le graphe **a** est un arbre. Les graphes **b** et **c** ne sont pas des arbres parce que **b** contient des cycles et que **c** n'est pas connexe

Le *degré* d'un sommet indique le nombre de branches qui lui sont connectées. Un sommet de degré 1 est appelé une *feuille*. Un *noeud* est un sommet *interne* associé à au moins deux branches et possédant un degré strictement supérieur à 1. Un arbre *binaire* (Fig.3) est un arbre dont tous les noeuds sont de degré strictement égal à 3.

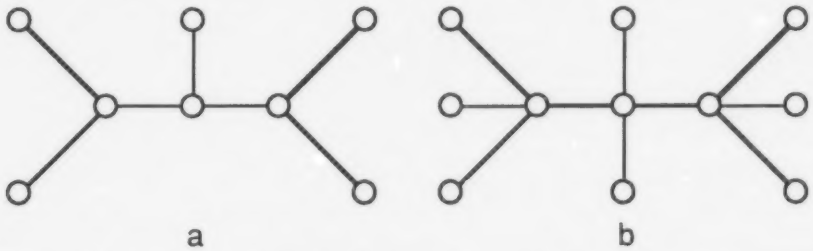


Fig. 3. L'arbre **a** est un arbre binaire alors que l'arbre **b** n'est pas binaire

Un arbre *étiqueté* (Fig. 4) est associé à un ensemble d'éléments ou d'objets fixés aux sommets de l'arbre. Tous les sommets peuvent être étiquetés ou uniquement les feuilles.

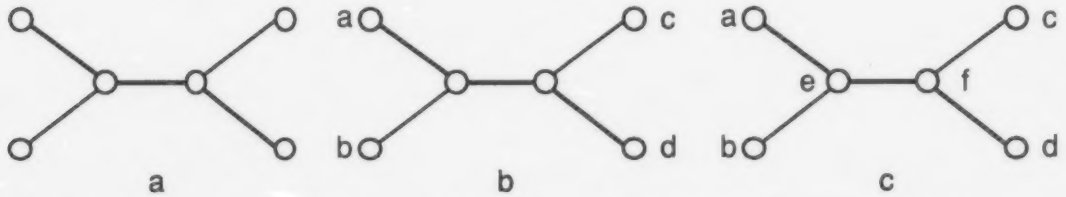


Fig.4. L'arbre **a** est non étiqueté. L'arbre **b** est étiqueté terminalement. L'arbre **c** est complètement étiqueté

Une arbre *planté* ou *enraciné* possède un sommet étiqueté comme la *racine* de l'arbre (Fig. 5). Cette racine induit une relation d'ordre sur les arêtes du graphe et implique une relation de "descendance" entre les sommets de l'arbre. La racine se trouve associée au noeud représentant l'ancêtre de tous les autres sommets; les feuilles sont considérées comme les descendants des sommets internes.

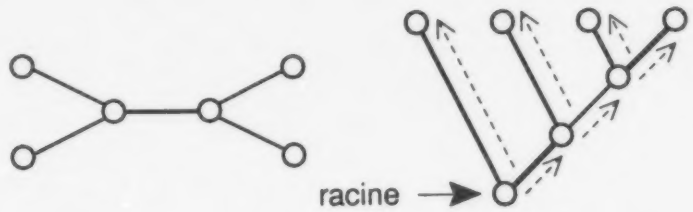
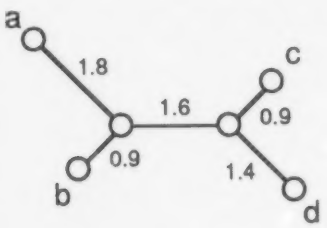


Fig. 5. L'arbre **a** est non enraciné alors que l'arbre **b** possède une racine induisant une direction aux arêtes.

Un arbre *valué* (Fig. 6) est associé à un ensemble de valeurs fixées à ses arêtes dans le but de représenter une fonction quelconque entre deux sommets adjacents. Un arbre *additif* est une arbre valué dont la distance entre deux sommets u et v est calculée en effectuant la somme de toutes les longueurs de branches se trouvant sur le chemin entre u et v. Chaque arbre additif est associé à une matrice de distances additives unique (Buneman, 1974).



	a	b	c
b	2.7		
c	4.3	3.4	
d	4.8	3.9	2.3

a b

Fig. 6. **a**) Un arbre additif, binaire, non enraciné, étiqueté terminalement et valué. **b**) la matrice additive correspondante.

Un *dendrogramme* est un cas particulier d'arbre additif enraciné où toutes les feuilles sont équidistantes de la racine (Fig. 7a). Les dendrogrammes sont aussi des arbres *sphériques* puisque leurs feuilles se trouvent situées sur la circonférence d'un cercle de rayon connu ayant comme centre la racine. Cette représentation implique une relation d'ordre entre les noeuds du dendrogramme en fonction de leur distance à la racine. Tout comme les arbres

additifs, les dendrogrammes sont associés à des matrices *cophénétiques* (Sokal et Rohlf 1962) représentant les distances entre chaque paires de feuilles (Fig. 7b) et à des matrices de *cardinalité* qui représentent la structure binaire de l'arbre (Fig. 7c).

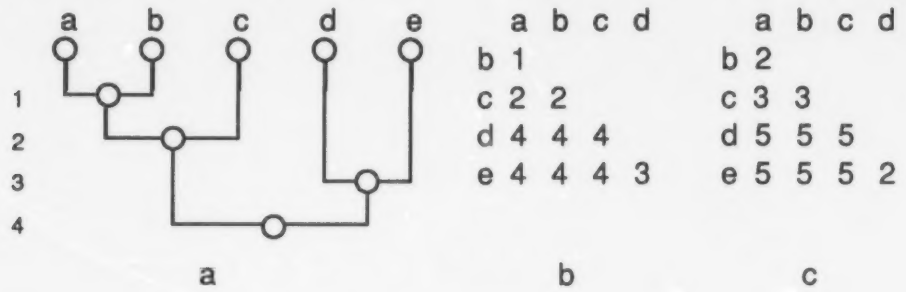


Fig. 7. **a**) Un dendrogramme **b**) la matrice cophénétique **c**) la matrice de cardinalité

Le concept d'arbre en biologie évolutive

Un arbre évolutif représente les relations phylogénétiques entre certaines espèces d'animaux ou de plantes. Cet arbre retrace l'histoire évolutive des lignées ancestrales ayant mené aux formes actuelles. Toute phylogénie est obligatoirement enracinée. La racine est une estimation de l'ancêtre primitif (hypothétique ou vrai dans le cas de fossiles). Seules les feuilles d'un arbre évolutif sont étiquetées et représentent un groupe d'espèces donné. On représentera la phylogénie par un arbre binaire non valué si l'on s'intéresse uniquement à la séquence de bifurcations de ses branches évolutives. Un arbre additif sera préféré dans les situations où l'on désire considérer les distances phylogénétiques interspécifiques. Lorsque les taux évolutifs des espèces sont identiques, un arbre phylogénétique pourra se présenter sous la forme d'un dendrogramme. Ces dendrogrammes ou classifications hiérarchiques sont fréquemment utilisés en biologie en raison de leurs propriétés mathématiques s'alliant aux concepts biologiques (Lapointe 1990).

Pourquoi comparer des arbres?

Un dendrogramme est une hypothèse évolutive représentant l'évolution d'un ensemble d'espèces. Nul ne peut vérifier cette hypothèse mais pourrait la *falsifier* par d'autres hypothèses qui contrediraient la première (Popper 1959). Il est en effet possible d'obtenir deux dendrogrammes totalement distincts en se basant sur différents jeux de caractères. La phylogénie du cerveau par exemple peut s'avérer très différente de celle du comportement. Le

premier dendrogramme pourrait ici permettre de falsifier le second et vice versa. L'intérêt de la comparaison de deux arbres évolutifs concurrents répond donc à la question visant à vérifier que les deux solutions sont effectivement différentes. Lorsque deux arbres sont déclarés similaires ou *congruents*, l'hypothèse évolutive sera *corroborée* (Penny, Foulds et Hendy 1982). Nous tenterons dans cette article de comparer certaines méthodes permettant d'évaluer statistiquement la congruence entre deux arbres. Nous nous pencherons plus spécifiquement sur la comparaison de dendrogrammes.

Comment comparer des arbres?

Diverses méthodes ont été décrites pour répondre à cette question. Nous retiendrons trois tests statistiques pouvant servir à la comparaison de classifications hiérarchiques. Comme nous l'avons vu auparavant (Lapointe 1990), un dendrogramme contient trois niveaux d'information:

a) une topologie

b) la position des étiquettes sur les feuilles de cette topologie

c) une échelle de ressemblance associée à la topologie

Les divers tests proposés se basent sur différents aspects de la définition mais évaluent tous la même hypothèse nulle générale stipulant que "les deux arbres comparés ne sont pas plus similaires entre eux que le seraient des paires d'arbres échantillonnés aléatoirement à partir de la même population statistique". La différence entre les tests de comparaison fait appel au concept d'arbre invoqué par les diverses méthodes. Le test de Mantel (1967) ne s'intéresse qu'à la position des étiquettes, Shao et Sokal (1986) considèrent en plus la topologie binaire des arbres alors que seule la méthode de Lapointe et Legendre (1990) tient compte des trois aspects de la définition caractérisant les dendrogrammes. Les tests que nous comparerons appartiennent cependant tous à la même famille des *tests par permutations*.

Qu'est-ce qu'un test par permutations?

La façon la plus simple de comparer deux arbres est de vérifier que ceux-ci sont plus ressemblant que toutes les paires d'arbres possibles possédant le même nombre de feuilles. Cette méthode idéale devient cependant incalculable lorsque le nombre d'arbres augmente en

fonction du nombre d'objets. On distingue par exemple 18 dendrogrammes différents comportant 4 feuilles alors que ce nombre est de 1 587 600 pour des arbres à 8 feuilles. La comparaison de tous les arbres possibles dans le cas de dendrogrammes à 8 feuilles serait de $1\ 587\ 600^2$, ce qui est impossible à calculer même pour les ordinateurs les plus rapides. Un test par permutations est une méthode statistique qui n'utilisera que certaines des 2.5×10^{12} comparaisons possibles en sélectionnant les paires d'arbres aléatoirement et équiprobablement. Ce type de test ne possède que très peu de conditions d'application. Il se base sur sa propre distribution, n'exige pas la normalité ni l'homoscédasticité des données, ni l'indépendance des valeurs, ne requiert aucune table statistique, et se calcule très facilement par ordinateur (Edgington 1987).

Trois étapes sont nécessaires pour l'évaluation statistique de la congruence entre deux arbres à l'aide d'un test par permutations: 1) On mesure premièrement la ressemblance entre nos deux arbres en utilisant un indice de consensus préalablement choisi. 2) On génère ensuite une distribution de référence de la statistique en répétant plusieurs fois la mesure de la ressemblance entre des arbres aléatoires. 3) Finalement, la signification statistique de la valeur de consensus de référence (celle mesurée entre nos deux arbres initiaux) est évaluée en se basant sur la distribution nulle ayant été construite à partir des comparaisons d'arbres aléatoires. La paire d'arbres originale sera déclarée congruente si la majorité des paires d'arbres aléatoires sont moins ressemblantes que les arbres initiaux.

Présentation des trois tests de comparaison

Nous avons ici affaire à trois tests par permutations qui permettent la comparaison de dendrogrammes. Notons que dans cette situation particulière, la comparaison de matrices cophénétiques peut se substituer à la comparaison de dendrogrammes.

Dans le cas du test de Mantel (1967), tel qu'appliqué par Hubert et Baker (1977) aux matrices cophénétiques, on s'intéresse à vérifier que les deux matrices sont plus similaires que des matrices dont on aurait permuté les lignes et les colonnes. Pour un dendrogramme, ceci consiste à étiqueter aléatoirement les feuilles de l'arbre. On se trouve donc à tester que nos deux arbres sont plus similaires quant à la position de leurs feuilles que des arbres topologiquement identiques avec d'autres étiquettes sur les feuilles.

La méthode de comparaison proposée par Shao et Sokal (1986) vise à vérifier que deux arbres binaires sont plus semblables entre eux que d'autres paires d'arbres binaires aléatoires. En fait, ce concept permet de tenir compte de la topologie ainsi que de la position des étiquettes. On néglige par contre ici la hauteur des branches en ne considérant que la séquence de bifurcations binaires des arêtes de l'arbre sans considération pour les longueurs de celles-ci.

La procédure statistique développée par Lapointe et Legendre (1990) permet d'évaluer la signification statistique de la ressemblance entre deux matrices cophénétiques représentant deux dendrogrammes indépendants. La méthode à doubles permutations génère une distribution de référence en comparant des arbres possédant une topologie binaire aléatoire, des feuilles étiquetées au hasard ainsi que des niveaux de fusion redistribués aléatoirement sur la topologie.

Comparaison théorique des trois méthodes

Les différences entre ces trois tests reposent sur la population de référence à partir de laquelle les arbres aléatoires sont sélectionnés. Dans chaque cas, l'échantillonnage est équiprobable mais l'effectif de la population ainsi que le nombre de comparaisons possibles varient selon le test utilisé. La population statistique du test de Mantel englobe l'ensemble des différentes manières d'étiqueter une topologie fixe (E). Le test de Shao et Sokal est représenté par la population statistique comportant tous les arbres binaires (B) distincts pour un nombre donné d'objets (Phipps 1975, Felsenstein 1978). Le test à doubles permutations développé par Lapointe et Legendre sélectionne des arbres aléatoires (D) parmi l'ensemble des dendrogrammes possibles (Frank et Svensson 1981). Les équations suivantes représentent respectivement le nombre de cas possibles pour chacun de ces tests.

Mantel (1967) $$E = n!/ 2^m \qquad \text{(éq.1)}$$

Shao et Sokal (1986) $$B = (2n-3)! / 2^{n-2}(n-2)! \qquad \text{(éq.2)}$$

Lapointe et Legendre (1990) $$D = n!\ (n-1)! / 2^{n-1} \qquad \text{(éq.3)}$$

Ces trois équations sont fonction du nombre d'objets n mais le test de Mantel (éq. 1) varie aussi selon la topologie de l'arbre où la variable m représente le nombre de noeuds possédant

exactement deux feuilles comme descendants (voir Fig. 8). On remarquera par exemple sur la figure 8 que 18 dendrogrammes sont possibles pour 4 objets. Parmi ceux-ci, 15 arbres binaires sont identifiables puisqu'on considère les arbres m et p, n et q, et o et r identiques lorsque la hauteur des niveaux de fusion est ignorée. Les étiquetages différents reconnus par Mantel sont fonction de la topologie. On distingue 12 cas pour la topologie A (m=1) alors que la topologie B (m=2) ne permet que 6 façons différentes d'étiqueter les feuilles.

Le tableau 1 nous indique que pour un nombre donné d'objets n>3, la relation suivante est toujours respectée indépendamment de la variable m:

$$En < Bn < Dn \qquad \text{(éq. 4)}$$

Ceci signifie qu'on retrouvera toujours plus de dendrogrammes que d'arbres binaires, non seulement car plus de topologies existent, mais aussi parce qu'il existe plus de manières d'étiqueter des dendrogrammes que de simples arbres binaires.

Tableau 1. Effectif des populations statistiques des tests de Mantel, Shao et Sokal, et Lapointe et Legendre en fonction du nombre d'objets n pour un m constant. Les valeurs E, B et D représentent respectivement le nombre d'étiquetages, d'arbres binaires et de dendrogrammes possibles.

n(m=1)	E	B	D
1	1	1	1
2	1	1	1
3	3	3	3
4	12	15	18
5	60	105	180
6	360	945	2700
7	2520	10395	56700
8	20169	135135	1587600
9	181440	2027025	57153600
10	1814400	34459425	257191200

Ces chiffres démontrent que les populations de référence des tests de Mantel et de Shao et Sokal représentent des sous-ensembles de la population totale de dendrogrammes considérée par la méthode de Lapointe et Legendre. On peut penser que les méthodes ne se basant pas sur l'entière population de dendrogrammes pour comparer des classifications hiérarchiques seront biaisées dans leurs conclusions statistiques. Cette perte d'information pourrait entraîner des résultats variables pour les trois méthodes.

Comparaison empirique des trois méthodes

La théorie démontre que les trois tests sous investigation sont fondamentalement

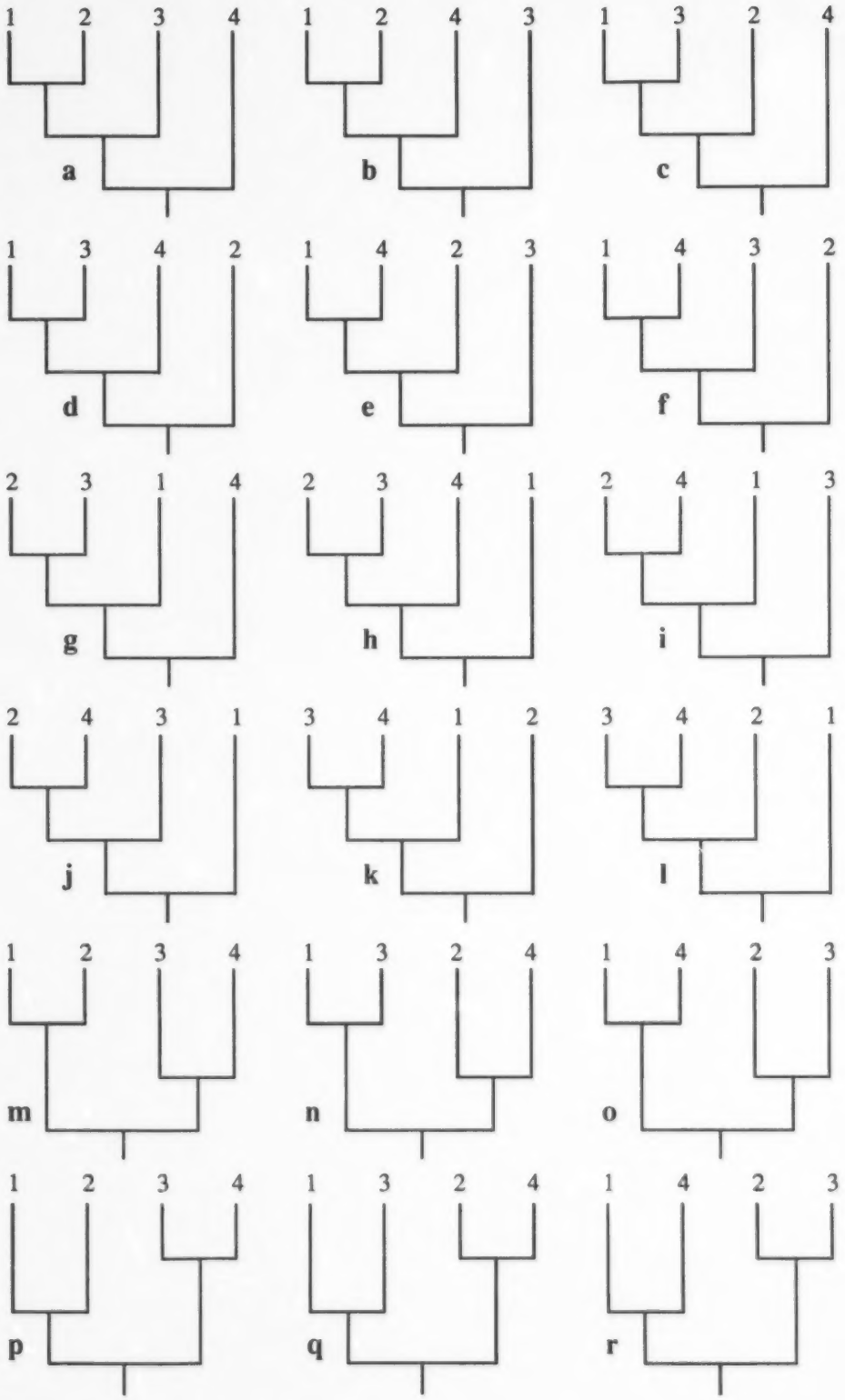


Fig.8. Représentation de tous les dendrogrammes possibles possédant quatre feuilles. Les arbres binaires distincts sont représentés par les dendrogrammes **a** à **o**. La topologie A (m=1) comprend les arbres **a** à **l** et la topologie B (m=2) est représentée par les arbres **m** à **r**.

différents. Nous tenons maintenant à confirmer les conclusions théoriques par des simulations visant à évaluer la performance relative des différentes méthodes. Nous viserons ici à comparer des paires de dendrogrammes à l'aide des trois tests décrits plus haut. Le but de ces simulations est double. Tout d'abord, nous voulons confirmer que le test à doubles permutations de Lapointe et Legendre se comporte tel que prévu lors des comparaisons de dendrogrammes. Nous désirons de plus déterminer la robustesse des deux autres méthodes. C'est-à-dire que nous voulons savoir si des tests non créés spécifiquement pour comparer les classifications hiérarchiques sont assez robustes pour s'appliquer à la comparaison de dendrogrammes. Ceci permettra de justifier ou de proscrire leur emploi.

Nous avons généré aléatoirement 500 paires de dendrogrammes portant cinq objets selon la procédure décrite par Lapointe et Legendre (sous presse). Chacune de ces paires fut comparée à l'aide des trois différentes méthodes pour un total de 1500 tests. Nous avons choisi le coefficient de corrélation linéaire de Pearson comme indice de ressemblance. Cette corrélation se calcule directement entre les matrices cophénétiques dans le cas des tests de Mantel et de Lapointe et Legendre. Pour la méthode de Shao et Sokal, la corrélation est mesurée sur les matrices de cardinalité associées aux dendrogrammes afin de ne considérer que la structure binaire des arbres. La signification statistique du test de Mantel fut évaluée en effectuant tous les permutations possibles des matrices cophénétiques. Dans le cas des deux autres méthodes, la distribution de référence fut reconstruite en sélectionnant 5000 paires d'arbres aléatoires. Le tableau 2 présente les résultats des simulations. On y indique, pour chaque test, le nombre de cas rejetant l'hypothèse nulle ainsi que la probabilité observée pour différentes valeurs critiques de signification.

Tableau 2. Résultats des simulations empiriques. Les valeurs présentées indiquent le nombre de cas où l'hypothèse nulle fut rejetée. La probabilité associée à chaque valeur est présenté entre parenthèse. Les résultats sont présentés pour trois différents seuils de signification alpha.

alpha	0,01	0,05	0,1
Mantel	0 (0,00)	5 (0,01)	20 (0,04)
Shao et Sokal	3 (0,006)	31 (0,062)	37 (0,074)
Lapointe et Legendre	5 (0,01)	25 (0,048)	49 (0,098)

On remarque tout d'abord que les probabilités observées pour le test à doubles permutations sont quasi-identiques aux probabilités théoriques. Ceci rencontre la théorie et justifie l'emploi de cette méthode pour comparer des dendrogrammes. En ce qui concerne les deux autres tests, il devient clair suite à l'examen des résultats que tous deux ont échoué leur épreuve de robustesse. Le test de Mantel semble trop conservateur en sous-estimant le nombre de cas où l'hypothèse nulle aurait dû être rejetée. Le test de Shao et Sokal démontre un comportement un peu plus chaotique. Il sous-estime la probabilité dans certaines situations et la surestime aussi quelquefois. On doit donc proscrire l'utilisation de cette méthode dans les cas où l'on désire comparer des dendrogrammes. Le test de Mantel, par contre, pourrait s'avérer utile si l'on recherche une méthode de comparaison pour dendrogrammes très conservatrice. Dans la majorité des cas où le test de Mantel fut significatif, le test de Lapointe et Legendre l'était aussi. On ne peut cependant généraliser cette loi car la comparaison d'arbres étiquetés de la même façon sur des topologies différentes entraînera des conclusions statistiques inverses pour ces ces deux méthodes.

Références

Barthélemy, J.-P., et A. Guénoche (1988), *Les Arbres et les Représentations des Proximités*, Méthode+Programme, Masson: Paris.

Buneman, P. (1974), "A Note on the Metric Properties of Trees," *Journal of Combinatorial Theory, Ser. B.*, 17: 48-50.

Edgington, E. S. (1987), *Randomization Tests, Second Edition*, Marcel Decker: New York.

Felsenstein, J. (1978), "The Number of Evolutionary Trees," *Systematic Zoology*, 27: 27-33

Frank, O. et K. Svensson. (1981), "On Probability Distributions of Single-Linkage Dendrograms," *Journal of Statistics and Computer Simulation*, 12: 121-131.

Hubert, L. J. et F. B. Baker. (1977), "The Comparison and Fitting of Given Classifications Schemes," *Journal of Mathematical Pscychology*, 16: 233-253.

Lapointe, F.-J. (1990), "La Comparaison Statistique de Classifications Hiérarchiques," pp: 213-222, dans *Recueil des textes des présentations du colloque sur les méthodes et domaines d'application de la statistique 1989*. Bureau de la statistique du Québec.

Lapointe, F.-J. et P. Legendre. (1990), "A Statistical Framework to Test the Consensus of Two Nested Classifications," *Systematic Zoology*, 39: 1-13.

Lapointe, F.-J. et P. Legendre. (sous presse), "On the Generation of Random Ultrametric Matrices for Comparing Dendrograms."

Mantel, N. (1967), "The Detection of Disease Clustering and a Generalized Regression Approach," *Cancer Research,* 27: 209-220.

Penny, D., L. R. Foulds, et D. M. Hendy. (1982), "Testing the Theory of Evolution by Comparing Phylogenetic Trees Constructed from Five Different Protein Sequences," *Nature,* 297: 197-200.

Phipps, J. B. (1975), "The Number of Classifications," *Canadian Journal of Botany,* 54: 686-688.

Popper, K. R. (1959), *The Logic of Scientific Discovery*, Hutchison: London.

Shao, K. et R. R. Sokal. (1986), "Significance Tests of Consensus Indices," *Systematic Zoology,* 35: 582-590.

Sokal, R. R. et F. J. Rohlf (1962), "The Comparison of Dendrograms by Objective Methods," *Taxon,* 9: 33-40.

ANALYSE DE COURBES DE CROISSANCE:

UN EXEMPLE EN CEPHALOMETRIE.

PAR

Litchy P. MELAGNE

Bernard CLEMENT

Dr Claude REMISE

Ecole Polytechnique

Université de Montréal

Sommaire

Cette étude a été conduite pour évaluer les effets du sexe et de l'âge sur les paramètres céphalométriques à partir d'une banque de données constituée au centre de croissance de l'Université de Montréal. Les sujets sont des enfants canadiens français des deux sexes âgés entre 9 et 16 ans.

L'analyse tend à mettre en évidence des courbes de croissance identiques ou différentes pour les garçons et les filles correspondant aux variables céphalométriques, et aussi des changements du taux et de la direction de croissance. Les résultats obtenus permettent des distinctions suivantes:

- 80% des mesures angulaires ne montrent aucune différence significative entre les sexes et présentent un taux de croissance nul ou stable dans le temps.
- Par contre, la majorité des mesures linéaires montrent des différences significatives entre les garçons et les filles avec des taux de croissance stables ou dépendant de l'âge. La direction de croissance varie selon le sexe.

Les résultats révèlent de façon générale une croissance forte chez les filles jusqu'à l'âge de 12.5 ans et un ralentissement après. Le phénomène inverse se produit chez les garçons.

1. Introduction

Depuis longtemps, l'analyse de courbes de croissance a été un problème important en biostatistique. Les deux problèmes sur lesquels les travaux ont été centrés concernent: l'estimation de courbes de croissance individuelles à partir de nombreux points de données, et l'estimation d'une courbe de croissance moyenne à partir d'un groupe d'individus.Le premier de ces problèmes a reçu beaucoup d'attention de plusieurs auteurs (Potthoff et Roy, 1964; Rao, 1965; Lee et Geisser, 1972, 1975). Cet article étend une approche récente qui cherche à combiner les données d'un groupe d'individus pour améliorer des paramètres individuels de croissance. Cette approche permet de tenir compte des observations manquantes qui arrivent fréquemment dans les études longitudinales, où les sujets sont observés à différents moments et/ou sous diverses conditions expérimentales (Kleinbaum, 1973; Rubin, 1976; Little et Rubin, 1987; Little, 1988). Les modèles de telles données combinent souvent des structures de covariance particulières avec des structures de la moyenne correspondant à la moyenne des mesures répétées d'une ou plusieurs variable(s) réponse(s). Selon les travaux de Helms (1984), Ware (1985), Jennrich et Schluchter (1986), on modélise la croissance à l'aide d'un polynôme fonction de l'âge du sujet en incorporant des covariables de sexe et d'interaction âge et sexe ainsi que la possibilité de modéliser la

matrice de covariance intra-sujet. Les estimateurs maximum de vraisemblance (MV) du modèle sont calculés par l'algorithme EM (Expectation-Maximisation) à partir d'un échantillon incomplet. L'estimation MV des paramètres du modèle exige des procédures d'ajustement itératives de par la nature même de l'algorithme. Les analyses sont exécutées à l'aide du programme 5V du progiciel BMDP. Ce pogramme permet l'ajustement des modèles aux données incomplètes de mesures répétées en utilisant les méthodes du maximum de vraisemblance ou du maximum de vraisemblance restreint. Il offre un grand choix de structures pour la matrice de covariance et permet de spécifier un modèle pour la fonction de croissance.

2. Description des données

2.1 Echantillon

Des céphalogrammes latéraux ont été pris annuellement chez deux cohortes de sujets au centre de recherche sur la croissance de l'Université de Montréal. Les sujets sont des enfants canadiens-français des deux sexes et sont sélectionnés au hasard d'une population d'enfants fréquentant le centre une fois par année pour des examens préventifs. Aucun d'eux n'a subi de traitements orthodontiques au préalable. Les données longitudinales couvrent les âges de 9 à 16 ans réparties sur deux échantillons de 40 filles et 25 garçons. Le nombre de céphalogrammes utilisés pour la cueillette des données céphalométriques est résumé dans le tableau ci-dessous.

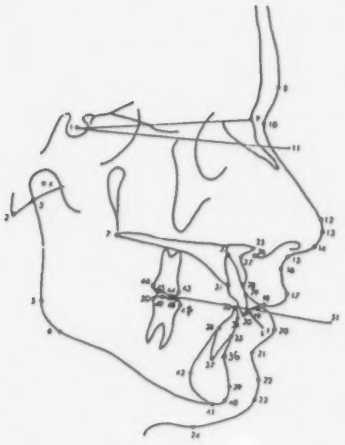
Age	nombre de céphalogrammes	Age	nombre de céphalogrammes
09	40	13	62
10	63	14	62
11	62	15	58
12	63	16	19

Tableau 1: Nombre de céphalogrammes utilisés.

2.2 Méthode d'enregistrement de données

L'enregistrement des données est fait par digitalisation sur chacun des céphalogrammes latéraux pris annuellement chez chaque sujet. La méthode employée est basée sur celle du Dr Charles J. Burstone de l'Université du Connecticut. Cette méthode comprend l'enregistrement de 52 points divers (tissus durs et tissus mous

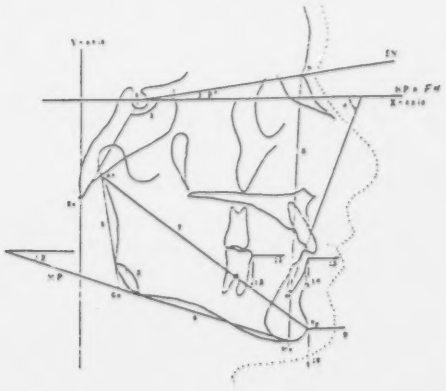
réunis) identifiés sur le céphalogramme (Fig.1) et à les trans-



Source [13]

Fig.1: Points céphalométriques

former en 61 variables céphalométriques (Fig.2) dans un espace géométrique où l'axe des abscisses est une horizontale (FH) obtenue en abaissant de 7 degrés le segment de droite S-N et l'axe des ordonnées est une perpendiculaire au plan FH qui se construit selon la position des points considérés. La digitalisation dans cette étude est l'oeuvre du Dr Robert Ledoux.



Source [13]

Fig.2: Mesures céphalométriques

Dans le but d'obtenir des données céphalométriques de qualité, un programme a été conçu pour transformer les 52 points en 61 variables ou mesures. Ce programme permet de calculer les angles et les distances correspondant aux mesures angulaires et linéaires.

A partir d'un céphalogramme, 61 variables sont définies et classifiées selon le tableau suivant. Notons que les variables 1 à 61 sont des mesures angulaires et linéaires, donc continues.

	angulaires	distances	total
Tissus durs	13	27	40
Tissus mous	5	16	21
Total	18	43	61

Tableau 2: Classification des variables.

2.3 Objectifs

Dans cette étude, on se propose de construire des courbes et tableaux de croissance pour chacune des variables et dégager les tendances de croissance générale. On doit définir quel modèle sera approprié pour décrire le comportement de chaque variable par rapport à l'âge et au sexe, et on va répondre aux questions suivantes pour chaque paramètre céphalométrique:

a) Quelle est la structure de dépendance intra-sujet?

b) Quelle est la courbe de croissance?

c) Les courbes de croissance sont-elles identiques pour chacun des deux sexes?

3. Méthodes d'analyse

3.1 Modèle général

Considérons la matrice de données suivante:

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1T} \\ y_{21} & y_{22} & \cdots & y_{2T} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ y_{T1} & y_{T2} & \cdots & y_{TT} \end{bmatrix}$$

où y_{ij} = valeur de la j^e variable du i^e sujet à un âge donné. Cette matrice de données n'est pas nécessairement complète pour toutes les variables. En effet, à cause de la présence des données manquantes et des mesures répétées dans le temps sur un même sujet, le modèle général proposé par Jennrich et Schluchter (1986) est:

$$Y \sim N(E(Y) , \Sigma) \qquad (1)$$

où $E(Y) = X\underset{\sim}{\beta}$ est la fonction de croissance et $\Sigma = \Sigma(\underset{\sim}{\theta})$ est une fonction d'un ensemble de q paramètres de covariance. Puisque les paramètres de régression $\underset{\sim}{\beta}$ se comportent indépendamment des paramètres de covariance $\underset{\sim}{\theta}$, le modèle (1) comporte deux parties distinctes: un modèle pour les valeurs moyennes impliquant les paramètres $\underset{\sim}{\beta}$, et un modèle pour la structure de covariance avec les paramètres $\underset{\sim}{\theta}$. De ce fait, la modélisation consistera à combiner divers choix de structure de moyenne et de covariance tels que proposés par Jennrich et Schluchter (1986), Waternaux, Laird et Ware (1989).

3.2 Fonction de croissance

On modélise la croissance à l'aide d'un polynôme fonction du temps en incorporant des covariables de sexe et d'interaction âge et sexe. La présence des termes d'interaction dans la matrice X a pour effet d'augmenter le nombre de paramètres de régression lorsque le degré de la fonction polynômiale croît. De plus, puisque les données s'échelonnent sur huit ans, nous avons arrêté notre choix sur des polynômes du 1er et 2^{e} degré.

3.3 Structures particulières de covariance

L'analyse des données longitudinales exigent la plupart du temps la modélisation de la matrice de covariance. Potthoff et Roy (1964) suggèrent des structures particulières dans l'analyse des données de courbes de croissance.

3.3.1 Matrice de covariance générale

Cette structure de covariance est utilisée dans les modèles de données incomplètes lorsqu'il n'existe aucune base théorique ou empirique de supposer une structure plus simple et plus facile à interpréter. Dans ce cas, la modélisation et l'estimation de Σ comportent $T(T+1)/2$ paramètres de covariance inconnus correspondant aux variances et covariances. Les paramètres sont établis comme suit:

$\theta_1 = \sigma_{11}$, $\theta_2 = \sigma_{21}$, $\theta_3 = \sigma_{22}$, $\theta_4 = \sigma_{31}$, . . . , $\theta_q = \sigma_{TT}$ où $q = T(T+1)/2$. Les σ_{ij} représentent les variances pour $i = j$ et les covariances pour $i \neq j$. Lorsque T est grand et les données sont incomplètes, l'utilisation de cette structure de covariance peut

mener à une perte d'efficacité dans l'estimation des paramètres de régression due au nombre élévé de paramètres de covariance à estimer (Ware,1985). En imposant des contraintes sur les paramètres de covariance de cette structure, le modèle se réduit à des structures alternatives plus simples et avec moins de paramètres.

3.3.2 Structure à bande (autorégressive générale)

Le nombre de paramètres dépend du nombre de mesures répétées. Lorsque les T réponses pour un sujet donné, $y_{i1},\ldots,y_{iT}$, sont ordonnées en fonction du temps,la structure des résidus est générée par un processus stationnaire où

$\theta_1 = \sigma^2 = \text{var}\,(y_{it})$, $\quad t = 1,\ldots,T$, et

$\theta_j = \text{cov}\,(y_{it}, y_{i,t-j})$, $\quad t = 2,\ldots,T$.

Par conséquent, le modèle de cette structure s'écrit comme suit :

$\Sigma = [\sigma_{jk}]$ où $\sigma_{jk} = \theta_r$, $r = |j - k| + 1$, $r = 1,\ldots,T$.

3.3.3 Modèle AR(1)

Le modèle de la structure AR(1) s'écrit:

$\Sigma = [\sigma_{jk}]$ où $\sigma_{jk} = \sigma_u^2\ \rho^{|j-k|} = \theta_1\theta_2$ $\quad j,k = 1,\ldots,T$

où θ_1 représente la variance de la réponse, i.e.,

$\theta_1 = \text{var}(y_{it})$ et

θ_2 la corrélation entre deux observations, i.e.,

$\theta_2 = \text{corr}\,(y_{it}, y_{i,t-j})$.

La corrélation tend à décroître et la variance reste stable lorsque le délai augmente, ce qui est conforme au comportement normal d'un processus autorégressif d'ordre un.

3.3.4 Structure d'équicorrélation (symétrie composée)

Cette structure qui suppose une variance constante et une corrélation commune entre les mesures correspond à un simple modèle de composantes de la variance avec une variance inter-sujet et une variance intra-sujet. Le modèle de cette structure peut s'écrire:

$$\Sigma(\underset{\sim}{\theta}) = \theta_1 U + \theta_2 I_T \quad \text{et} \quad \underset{\sim}{\theta} = (\theta_1, \theta_2)$$

où θ_1 et θ_2 représentent les variances inter et intra sujets respectivement, U est une matrice (T x T) composée de 1 et I_T est la matrice identité de même dimension que U. La matrice de covariance correspondant à cette structure possède des éléments $\theta_1+\theta_2$ sur la diagonale et θ_2 ailleurs. La structure d'équicorrélation est utile dans les modèles de courbes de croissance parallèles.

Le tableau ci-dessous donne un résumé des structures décrites.

Structure	nombre de paramètres	Description
générale	T(T+1)/2	$\sigma_{11}=\theta_1$, $\sigma_{12}=\theta_2$, ..., $\sigma_{TT}=\theta_q$
à bande	T	$\sigma_{ij}=\theta_k$, $k = \lvert i-j \rvert + 1$
ar(1)	2	$\sigma_{ij}=\sigma^2 \rho^{\lvert i-j \rvert}$
		$\theta_1 + \theta_2$ $\quad i = j$
équicorrélation	2	$\sigma_{ij}=$
		θ_1 $\quad i \quad j$

Tableau 3: Structures pour la matrice de covariance.

3.4 Estimation

L'estimation des paramètres de covariance avec la méthode vraisemblance maximale s'effectue de façon itérative en utilisant plusieurs algorithmes (NR=Newton-Raphson, FS=Fisher Scoring et GEM=Generalized EM). Si $\hat{\underset{\sim}{\theta}}$ est le vecteur des estimateurs MV des paramètres de covariance et $\hat{\Sigma} = \Sigma(\hat{\underset{\sim}{\theta}})$ pour i = 1,...,n, les estimateurs de $\underset{\sim}{\beta}$ sont les estimateurs moindres carrés pondérés:

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i' \hat{\Sigma}_i^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i' \hat{\Sigma}_i^{-1} y_i\right)$$

Pour différencier la courbe de croissance entre les deux groupes, la statistique khi-deux de Wald permettant de tester l'hypothèse H_0:

$$L\,\underset{\sim}{\beta} = 0$$

est donnée par

$$W = (L\underset{\sim}{\beta})'(LV\,L')^{-1}(L\underset{\sim}{\beta})$$

où $V = \sum_{i=1}^{n}(X_i \Sigma^{-1} X_i)^{-1}$ et L est une matrice (m x p) de constantes, m étant égal au nombre d'éléments du vecteur $\underset{\sim}{\beta}$. Sous H_0, cette statistique est distribuée comme une khi-deux avec m degrés de liberté. On rejette H_0 si $W > \chi^2_{\alpha}(m)$ où α est un seuil.

4. Résultats

4.1 Structures particulières de covariance

Le tableau 3 concerne la fréquence de chaque structure de covariance dans l'analyse. Il indique que la structure d'équicorrélation apparaît plus souvent,surtout avec les tissus mous. Elle est suivie curieusement par la covariance générale. Notons la faible performance des structures autorégressives, en particulier le modèle AR(1) qui n'apparaît pas dans les tissus mous. On constate une nette distinction des structures de dépendance entre les tissus d'après les résultats du tableau.

Mesures	Tissus	Générale	A bande	Symétrie	AR(1)	Total
Linéaires	Durs	10	4	1	3	18
	Mous	1	0	14	0	15
Angulaires	Durs	9	7	2	3	21
	Mous	0	1	4	0	5
Total		20	12	21	6	59

Tableau 3: Fréquence des structures de covariance.

4.2 Fonction de croissance

Le tableau ci-dessous résume les différents types de croissance correspondant à chaque paramètre céphalométrique. On voit, d'après les résultats, que la croissance linéaire permet de décrire le comportement de plusieurs variables avec une vitesse

de croissance stable dans le temps. Notons aussi que ce type de croissance présente des courbes identiques pour la majorité des mesures angulaires.On constate par ailleurs que plusieurs mesures de distances (34) sont décrites par une croissance linéaire. On voit une presque égalité entre les deux types de croissance. Des différences plus marquées se rencontrent au niveau de la croissance quadratique. Les graphiques C.5 et C.7 sont des exemples de croissance obtenue.

	Mesures		
	Angulaires	distances	Total
Horizontale	1	1	2
Linéaire			34
identique	13	9	
différente	1	11	
Quadratique			23
identique	1	2	
différente	2	18	

Tableau 4: Types de croissance.

L'analyse a mis en évidence deux distinctions majeures:

1. Des paramètres céphalométriques qui ne montrent pas de différence significative entre les sexes: les courbes de croissance sont identiques et décrivent le même comportement tant pour les garçons que pour les filles. 83% des angles et 30% des distances présentent ce type de comportement.

2. D'autres paramètres, par contre, présentent des différences significatives entre les sexes. Pour les paramètres avec une croissance quadratique, on observe une poussée de croissance rapide chez les filles entre 9 et 12 ans suivie d'une décéléra-

tion à partir de la 13e année. Mais on constate le phénomène inverse chez les garçons.

Les résultats obtenus dans cette étude confirment ceux obtenus dans d'autres études (Bishara, 1981, 1985; Jamieson, 1982; Mitani, 1977).

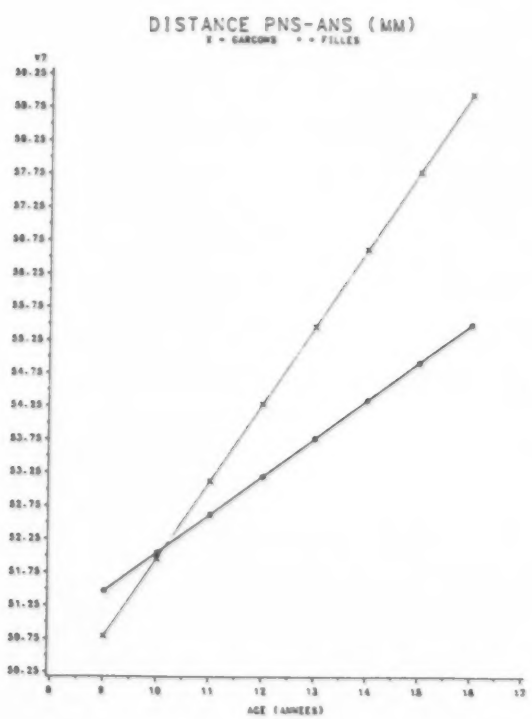


Figure C.5: Graphique de V7 vs âge

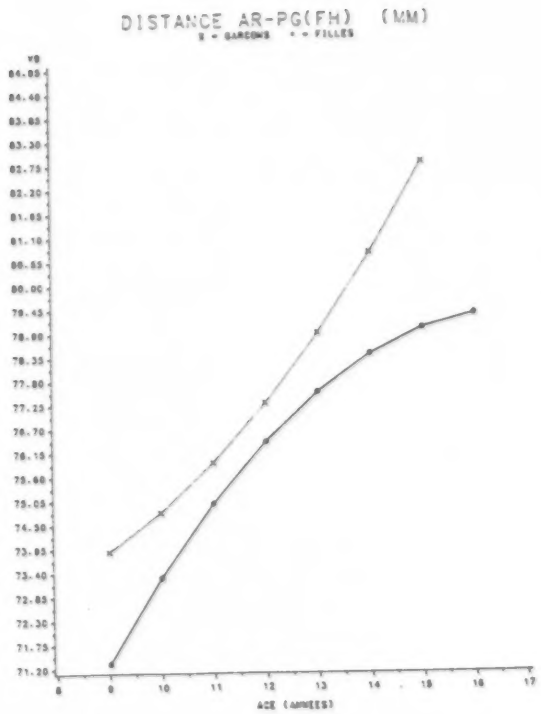


Figure C.7: Graphique de V9 vs âge

5. Discussion

L'approche importante utilisée dans cette étude pour analyser les mesures répétées avec données manquantes consiste à modéliser la matrice de covariance et la valeur moyenne. La capacité de modéliser la matrice de covariance mène à des estimations efficaces de β, surtout lorsque les données comportent beaucoup d'observations manquantes. Les comparaisons des algorithmes utilisés (NR, FS,GEM) pour estimer les paramètres de covariance sont difficiles à faire puisque cela dépend du degré d'efficacité de ces algorithmes. L'algorithme NR converge généralement avec un nombre faible d'itérations à un coût élévé par itération. A l'autre extrême, l'algorithme GEM, à un coût minime, exige un nombre d'itérations élévé. Lorsque le nombre de paramètres de covariance q est faible, n'importe lequel de ces algorithmes est généralement satisfaisant mais on préfère l'algorithme NR à cause de sa convergence rapide.Avec q est élévé,i.e.,le nombre T de mesures répétées est supérieur à dix, seulement GEM est possible à cause du coût minime par itération.

Dans la fonction de croissance,on a introduit un terme d'interaction des covariables sexe et âge afin de mesurer les différences entre les garçons et les filles. Lorsqu'elle est importante, les

courbes de croissance obtenues s'interceptent. Autrement, on se retrouve avec des courbes presque parallèles, ce qui donne des effets additifs.

Enfin,le programme 5V de BMDP est très efficace pour l'analyse de mesures répétées avec données manquantes car il permet un grand choix de structures de covariance. Le seul désavantage de ce programme est qu'il ne fait l'analyse des résidus. Toutefois, à cause du nombre important de variables à étudier, quelques analyses ont été effectuées pour assurer la validité des résultats. Finalement, les résultats de cette étude fournissent une description préliminaire raisonnable de la croissance craniofaciale et des tissus mous. Les implications cliniques de ces résultats sont évidentes puisqu'elles donnent des descriptions plus précises disponibles sur la croissance des paramètres céphalométriques pour les filles et les garçons de souche canadienne-française.

Bibliographie

[1] Bishara et al: Longitudinal changes in standing height and mandibular parameters between the ages of 8 to 17 years. Am J Orthod, vol.80 No.2:115-135, 1981.

[2] Bishara S.E. et Jacobsen J.R: Longitudinal changes in three normal faces types. Am J Orthod, 88:466-502, 1985.

[3] Jamieson et al: Longitudinal changes in the maxilla and the maxillary-mandibular relationship between 8 and 17 years of age. Am J Orthod, vol.82 No.3:217-230, 1982.

[3] Jennrich Robert I. et Schluchter Mark D.: Unbalanced repeated measures models with structured covariances matrices. Biometrics, 42:805-820, 1986.

[4] Kleinbaum D.G.: Testing linear hypotheses in generalized multivariate linear models. Comminications in Statistics, 1:433-457, 1973.

[5] Koziol James A.: An introduction to multivariate growth curve analysis. Growth, 50:259-272, 1986.

[6] Lee J. et Geisser S.: Growth curve prediction. Sankhya, Series A 34, 393-412, 1972.

[7] Lee J. et Geisser S.: Applications of growth curve prediction. Sankhya Series A 37, 239-256, 1975.

[8] LaVange L.M. et Helms R.W.: The analysis of incomplete longitudinal data with modeled covariance structures. University of North Carolina Instistute of Statistics Mimeo Series, No. 1449, 1983.

[9] Little R.J.A.: Robust estimation of the mean and covariance matrice from data with missing values. Applied Statistics, Serie A, 37:23-39, 1988.

[10] Little R.J.A. et Rubin D.B.: Statistical analysis with missing values. Wiley, New-York, 1987.

[11] Mémoire de maîtrise par Litchy P. Melagne
Ecole Polytechnique, Montréal, 1990.

[12] Mitani Hidéo: Prepubertal growth of mandibular prognathism. Am J Orthod, vol.80 No.5:546-553, 1981.

[13] Notes de cours de céphalométrie par le Dr Jack Turckewicz.

[14] Potthoff R.F. and Roy S.N.: A generalised multivariate analysis of variance model especially for growth curve problems. Biometrika, 51:313-326, 1964.

[15] Rao C.R.: The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. Biometrika, 52:447-458, 1965

[16] Rochon James and Helms Ronald W.: Maximum likelihood estimation for incomplete repeated measures experiments under an ARMA covariance structure. Biometrics, 45:207-218, 1989.

[17] Rubin D.B.: Inferences et missing data. Biometrika, 63:581-590, 1976.

[18] Rubin D.B.: Characterizing the estimation of parametres in incomplete data problems. Journal of Amer. Stat. Assoc., 69:467-474, 1974.

[19] Schluchter Mark D.: Analysis of incomplete multivariate data using linear models with structured covariances matrices. Statistics in Medecine, 7:317-324, 1988.

[20] Ware J.H: Linear models for the analysis of longitudinal studies. American Statistical Association, volume 2, 39:95-101, 1985.

Etude des données influentes sous le modèle de régression de Cox

Daniel Morissette
Département de mathématiques et de statistique
Université Laval, Sainte-Foy (Québec)

Le modèle de régression de Cox ("proportional hazards model") est très souvent employé dans l'étude des durées de vie lorsque des observations sont censurées. Son attrait principal vient du fait que ce modèle regroupe plusieurs familles de lois dont la loi de Weibull et la loi exponentielle. Dans ce qui suit nous présenterons quelques approches visant à détecter les individus les plus influents lorsque ce modèle est utilisé. Une attention particulière sera portée à l'approche développée par Pettitt et Bin Daud, approche inspirée des idées de Cook: l'influence locale. Pour terminer, un exemple d'application sera présenté.

1. Introduction

Considérons la variable aléatoire continue $T > 0$ représentant la durée de vie des individus de la population étudiée. On prélève de cette population un échantillon de taille n et pour chacun des individus on mesure t_i, la durée de vie ou le temps de censure de l'individu i et $\mathbf{x}_i = (x_{1i} , \cdots , x_{pi})$, le vecteur de variables explicatives de l'individu i. Par ailleurs, on notera $\boldsymbol{\beta}' = (\beta_1 , \cdots , \beta_p)$, le vecteur des coefficients de la régression.

Le modèle de régression de Cox spécifie que la relation entre le taux de panne des individus et le vecteur de variables explicatives est donnée (Cox, 1972) par

$$h(t \mid \mathbf{x}) = h_0(t) \exp(\mathbf{x}\boldsymbol{\beta}),$$

où $h_0(t) = h(t \mid \mathbf{x} = 0)$ est le taux de panne de base, i.e. le taux de panne des individus ayant le vecteur de variables explicatives nul. L'estimation de $\boldsymbol{\beta}$ s'obtient en maximisant à l'aide d'une procédure itérative la fonction de vraisemblance partielle

$$L(\boldsymbol{\beta}) = \prod_{i \in D} \frac{\exp(\mathbf{x}_i \boldsymbol{\beta})}{\sum_{l \in R_i} \exp(\mathbf{x}_l \boldsymbol{\beta})}, \qquad (1)$$

où D est l'ensemble des individus décédés (i.e. les individus qui sont non-censurés) et R_i est l'ensemble des individus à risque au temps t_i, i.e. l'ensemble des individus encore vivants et non-censurés juste avant le temps t_i. Remarquons qu'une procédure itérative sera nécessaire pour résoudre cette équation. Or, l'estimation de β obtenue sera parfois grandement influencée par quelques individus. Dans la section qui suit on présente des approches visant justement à déterminer ces individus influents.

2. Détection des individus influents

Soient $\hat{\beta}$, l'estimation de β obtenue en maximisant la fonction de vraisemblance partielle (1) et $\hat{\beta}_{(k)}$, l'estimation de β lorsque l'individu k ne fait plus partie de l'échantillon. L'influence de l'individu k sur l'estimation $\hat{\beta}$ peut être définie par

$$\hat{\beta} - \hat{\beta}_{(k)}.$$

Il est alors possible de calculer exactement l'influence de chacun des individus. Par exemple, pour obtenir l'influence de l'individu k il suffira d'exclure l'observation k et de modéliser les $n - 1$ observations restantes de l'échantillon. Il va sans dire cependant qu'une telle procédure, répétée pour chaque individu, demande un temps informatique important compte tenu du fait que l'estimation de β se fait au moyen d'une procédure itérative. Pour cette raison, on cherchera une autre approche à ce problème.

2.1 Influence des individus

Une façon possible de déterminer l'influence des individus serait de trouver une forme explicite approchant le vecteur $\hat{\beta} - \hat{\beta}_{(k)}$. C'est cette approche que Cain et Lange (1984), Reid et Crépeau (1985) de même que Storer et Crowley (1985) ont utilisée. Les composantes de ces approximations fournissent pour chaque individu k une mesure de son influence. Les observations les plus grandes en valeur absolue indiquent alors les individus les plus influents. Or, l'inconvénient de cette approche est qu'à chaque individu on associe une mesure d'influence pour chaque variable explicative du modèle (c'est donc dire qu'il s'agit d'un vecteur à p composantes). Ce problème implique qu'on devra soit combiner les composantes de ce vecteur afin d'obtenir une mesure unique de l'influence de chacun des individus ou encore se limiter à l'étude des observations influentes composante par composante (i.e. variable par variable). Dans cette dernière situation on pourra par exemple construire les p

graphiques de $| \hat{\beta}_j - \hat{\beta}_{j(k)} |$ en fonction de la variable explicative x_j $(j = 1, \cdots, p)$.

2.2 Influence locale

Une façon plus générale de détecter les individus influents a été proposée par Cook (1986). Il introduisit à cet effet $\omega' = (\omega_1, \cdots, \omega_n)$ le vecteur de poids associé aux individus et aussi le *déplacement de la log-vraisemblance* ("likelihood displacement") définie comme étant

$$2 [\log L(\hat{\beta}) - \log L(\hat{\beta}(\omega))], \tag{2}$$

où $\hat{\beta}(\omega)$ maximise la fonction de vraisemblance partielle pondérée fournie par Cain et Lange (1984):

$$L(\beta | \omega) = \prod_{i \in D} \frac{\exp(x_i \beta \omega_i)}{\left[\sum_{l \in R_i} \omega_l \exp(x_l \beta) \right]^{\omega_i}}. \tag{3}$$

On remarque que si le poids du $k^{ème}$ individu est nul ($\omega_k = 0$) et que le poids de tous les autres individus est égal à 1 alors la fonction de vraisemblance partielle pondérée $L(\beta | \omega)$ sera équivalente à la fonction de vraisemblance partielle $L(\beta)$ lorsque l'individu k ne fait plus partie de l'échantillon et $\hat{\beta}(\omega)$ sera équivalente à $\hat{\beta}_{(k)}$. On s'intéressera alors à trouver la direction dans laquelle le déplacement de la log-vraisemblance est la plus importante. En d'autres termes on cherchera le vecteur de poids ω qui maximise l'expression (2). Pour ce faire on notera $\omega_0' = (1, \cdots, 1)$, le vecteur donnant un poids unitaire à chaque individu, $A(\hat{\beta})^{-1}$, la matrice $(p \times p)$ de variances-covariances de $\hat{\beta}$ et enfin $\Delta = (\Delta_1, \cdots, \Delta_n)$, la matrice $(p \times n)$ dont la $k^{ème}$ colonne est donnée par

$$\Delta_k = \left[\frac{\partial}{\partial \omega_k} U(\beta | \omega) \right]_{\beta, \omega_k = 1},$$

où

$$U(\beta | \omega) = \frac{\partial \log L(\beta | \omega)}{\partial \beta}.$$

On remarque maintenant que si à tous les individus on attribue le même poids, i.e. $\omega = \omega_0$, alors l'estimation de β obtenue de la fonction de vraisemblance partielle pondérée (3) sera identique à celle qu'on obtiendrait de la fonction de vraisemblance partielle (1); bref, on constate que $\hat{\beta}(\omega_0) = \hat{\beta}$. On notera aussi $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_n$ les

valeurs propres de la matrice $\Delta' \mathbf{A}(\hat{\boldsymbol{\beta}})^{-1} \Delta$ $(n \times n)$ et $\mathbf{l}_1, \mathbf{l}_2, \cdots, \mathbf{l}_n$, des vecteurs propres de norme unitaire associés à ces valeurs propres. Pettitt et Bin Daud (1989) ont obtenu une approximation de (2) dans un voisinage de ω_0 (d'où vient le nom d'influence locale) soit

$$2 [\log L(\hat{\boldsymbol{\beta}}) - \log L(\hat{\boldsymbol{\beta}}(\omega))] \approx (\omega_0 - \omega)' \Delta' \mathbf{A}(\hat{\boldsymbol{\beta}})^{-1} \Delta (\omega_0 - \omega). \qquad (4)$$

Mentionnons au passage que cette approximation nous évite d'avoir recours à une procédure itérative. Or, l'équation (4) permet alors de conclure que le déplacement de la log-vraisemblance (2) sera approximativement maximisé et prendra la valeur γ_1 lorsque le vecteur de poids des individus ω sera tel que $(\omega_0 - \omega)$ égalera $\pm \mathbf{l}_1$, où, rappelons le, $\omega_0' = (1, \cdots, 1)$. Les individus dont l'influence locale sera grande seront ceux qui auront un poids ω_k petit (près de zéro) ou de façon équivalente, puisque le vecteur $\mathbf{l}_1$ est de norme unitaire, ceux qui correspondront aux composantes du vecteur propre $\mathbf{l}_1$ grandes en valeur absolue (près de 1). Cook suggéra également que si la valeur propre γ_1 est supérieure à 1 alors l'influence locale observée est importante.

De ces résultats on peut déduire une procédure graphique qui permettra de détecter les individus les plus influents (au sens de l'influence locale de Cook). Pour chaque valeur propre γ_j plus grande que 1, on portera en graphique les valeurs absolues des composantes du vecteur $\mathbf{l}_j$ en fonction des durées de vie observées (ou des censures). On retrouvera donc sur ces graphiques les couples $(| l_{ji} | , t_i)$, où t_i est la durée de vie observée (ou le temps de censure) du $i^{ème}$ individu et l_{ji} est la $i^{ème}$ composante du vecteur propre $\mathbf{l}_j$ associé à la valeur propre γ_j. Sur chacun de ces graphiques les observations les plus élevées correspondront aux individus les plus influents.

3. Exemple d'application

Pour illustrer nos propos, nous présenterons un exemple d'application sur un jeu de données tiré de Lawless (1982, p. 6). On y trouve les temps de survie (ou de censure) et aussi la mesure de 7 variables explicatives observées sur 40 individus souffrant d'un cancer du poumon. Les buts visés par cette recherche sont de quantifier l'effet

1) de deux traitements de chimiothérapie (ancien, nouveau) et
2) de quelques facteurs tels la condition physique générale du patient, son âge, etc.

sur le temps de survie des patients.

Tableau 1: Seuil observé des tests pour 2 variables

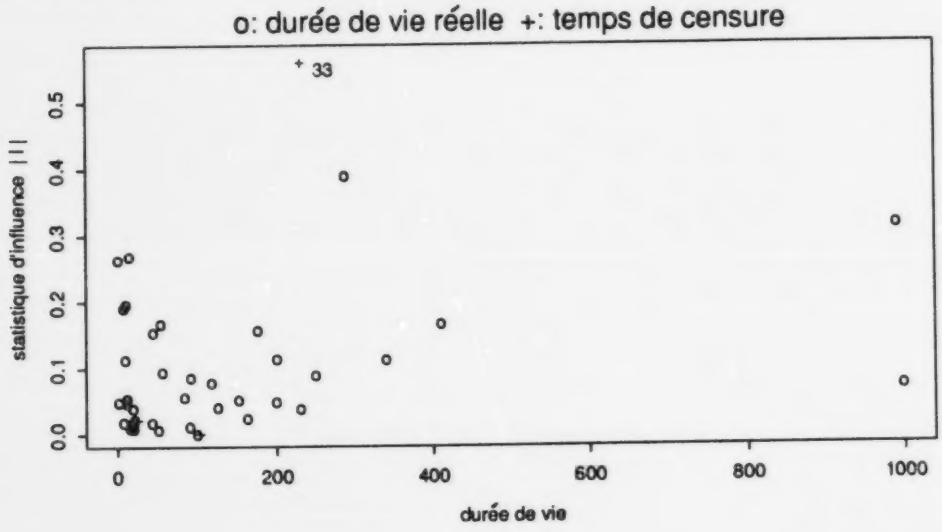
$H_0 : \beta_{CONDITION} = 0$	$H_0 : \beta_{TYPE} = 0$
.00 **	.08 *

*: significatif au seuil 10 %

**: significatif au seuil 1 %

L'analyse statistique sous le modèle de régression de Cox tend à montrer que parmi les 7 variables explicatives, seulement 2 ont un effet significatif sur le temps de survie des patients. Ces variables sont la condition physique générale du patient (CONDITION) et la présence ou l'absence du type de tumeur "adeno" (TYPE). Le tableau 1 présente le seuil observé ("p-value") des tests pour ces 2 variables.

Figure 1: valeur absolue de l_1 en fonction de la durée de vie ($\gamma_1 = 1.64$)



Suite à cette analyse, nous avons appliqué la procédure graphique suggérée à la section 2.2 afin de détecter les individus influents (au sens de l'influence locale). Les graphiques obtenus sont présentés aux figures 1, 2 et 3. La figure 1, par exemple, présente le graphique des composantes de l_1 en valeur absolue, l_1 étant le vecteur propre associé à la plus grande valeur propre γ_1, en fonction du temps de survie observé (ou de censure) des patients. On constate d'une part que $\gamma_1 = 1.64$ est plus

Figure 2: valeur absolue de l_2 en fonction de la durée de vie ($\gamma_2 = 1.11$)

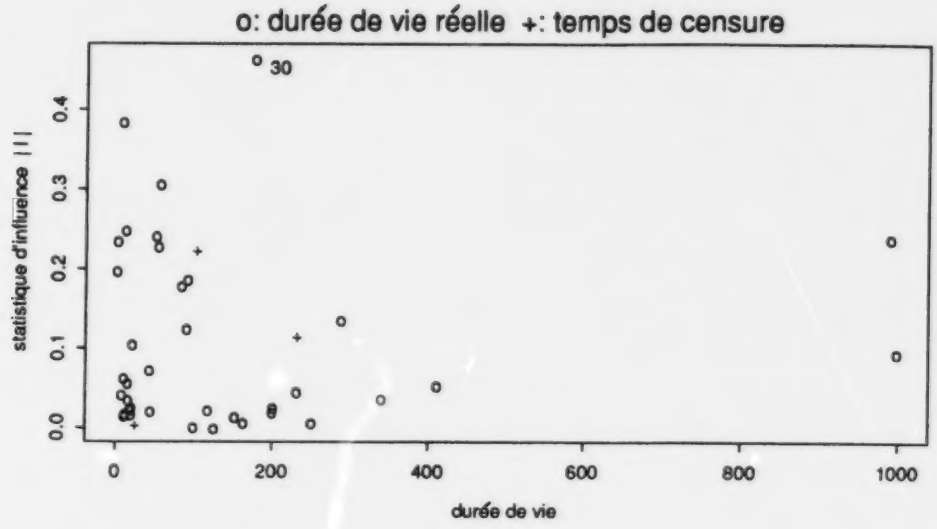
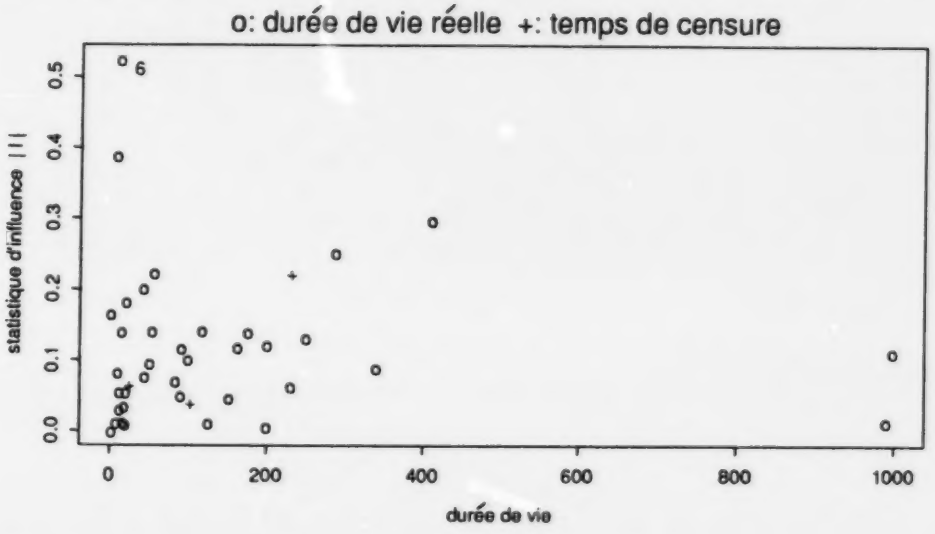


Figure 3: valeur absolue de l_3 en fonction de la durée de vie ($\gamma_3 = 1.01$)



grande que 1 laissant croire que l'influence locale observée est importante et d'autre part que l'observation # 33 est celle qui a la plus grande influence dans cette direction. De la même façon, les observations # 30 et # 6 sont mises en évidence par les figures 2 et 3 respectivement. Ainsi, pour vérifier l'influence réelle sur l'estimation du vecteur des coefficients β, on calculera à nouveau l'estimation de β mais cette fois en enlevant une à une ces observations. Les résultats des tests sur les échantillons amputés sont

Tableau 2: Seuil observé des tests pour 2 variables (avec 1 observation en moins)

# observation enlevée	$H_0 : \beta_{CONDITION} = 0$	$H_0 : \beta_{TYPE} = 0$
33	.00 **	.07 *
30	.00 **	.21
6	.00 **	.09 *

*: significatif au seuil 10 %

**: significatif au seuil 1 %

présentés au tableau 2. Les seuils observés des tests, comparés à ceux du tableau 1, ne varient pas de façon importante lorsque l'observation # 33 ou # 6 ne fait plus partie de l'échantillon alors que pour l'observation # 30, la variable TYPE n'a plus un effet significatif sur le temps de survie des patients. Le seuil observé du test passe en effet de .08 à .21 lorsque cette observation est enlevée de l'échantillon. On peut donc conclure que l'individu # 30 à une influence très importante sur l'estimation de β et plus précisément sur le coefficient de la régression β_{TYPE}.

4. Conclusion

Il est facile de trouver des applications dans lesquelles la présence d'observations influentes peut conduire à des conclusions erronées et entraîner des conséquences parfois graves. Pour ces raisons et par souci de toujours tirer les conclusions les plus justes l'étude des données influentes devrait, au même titre que l'étape de validation du modèle par exemple, faire partie intégrante de toute analyse statistique rigoureuse. L'analyse sous le modèle de régression de Cox ne fait pas exception à cette règle et l'influence locale de Cook telle qu'appliquée à ce modèle pourra nous être d'une grande utilité dans l'exécution de cette tâche.

BIBLIOGRAPHIE

Cain, K.C. et Lange, N.T. (1984). Approximate case influence for the proportional hazards regression model with censored data. *Biometrics*, **40**, 493-499.

Cook, R.D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society*, B, **48**, 133-169.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, B, **34**, 187-220.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. New-York: Wiley.

Pettitt, A.N. et Bin Daud, I. (1989). Case-weighted measures of influence for proportional hazards regression. *Applied Statistics*, **38**, 51-67.

Reid, N. et Crépeau, H. (1985). Influence functions for proportional hazards regression. *Biometrika*, **72**, 1-9.

Storer, B.E. et Crowley, J. (1985). A diagnostic for Cox regression and general conditional likelihoods. *Journal of the American Statistical Association*, **80**, 139-147.

UTILISATION COMPLÉMENTAIRE DE L'ANALYSE DES DONNÉES (AFC,ACP) ET DES ANALYSES PARAMÉTRIQUES EN SANTÉ ET EN SÉCURITÉ DU TRAVAIL

T. Petitjean Roget[1], A. Arsenault[2]

Résumé

La recherche en santé et en sécurité du travail, domaine relativement neuf, se prête peu à une planification expérimentale lorsqu'il s'agit de valider des hypothèses concernant des incidents-accidents. On utilisera plutôt des techniques de modélisation de l'univers des impacts (physiologiques, économiques, etc.), de l'univers des attributs (Descripteur de groupes, d'événements, etc.) et de leur interaction. Ces hypothèses à posteriori pourront faire l'objet d'analyses paramétriques.

[1]Institut de recherche en santé et en sécurité du travail

[2]Institut de cardiologie de Montréal

Introduction

Si la statistique fut déjà considéré comme l'art de faire mentir les chiffres, l'informatique scientifique est souvent confondue avec le département des miracles par les chercheurs en santé et en sécurité du travail: "J'ai reçu quelques critiques concernant mon rapport final, qu'en penses-tu?".

L'utilisation des techniques graphiques développées par l'école française d'analyse de données permet de mieux visualiser l'information disponible, de dégager des hypothèses qui seront explorées par des analyses paramétriques. Cette approche sera illustrée par un ensemble de données comportant des mesures physiologiques cardiovasculaires et des variables nominales portant sur des évènements stresseurs récents et des attitudes.

L'hypothèse générale cherche à identifier les évènements et attitudes ayant un impact potentiel sur la physiologie du système cardiovasculaire.

Position du problème

D'une façon générale, en santé et en sécurité du travail, on doit modéliser deux univers:

a) L'univers des impacts: physiologiques, économiques, etc.

b) L'univers des attributs "potentiellement explicatifs": descripteurs de groupe, paramètres de contrôle, etc.

On doit aussi modéliser les interactions entres ces deux univers.

L'exemple qui servira à illustrer cette problématique est extrait d'une étude du stress en milieu hospitalier (Arsenault, Dolan, 1984) et concernant les 179 individus retenus pour les fins de l'étude. Nous disposons:

1. d'un univers des conséquences comprenant:

a) des mesures de tension artérielle systolique et diastolique et de fréquence cardiaque.

b) répétées trois fois (triple lecture).

c) en position debout, couchée et à nouveau debout.

2. d'un univers potentiellement associés à des réactions structurelles dans l'univers des conséquences:

a) 35 événements intercurrents de la vie, côtés positifs ou négatifs, répartis en quatre univers(Famille, personnel, travail, finance).

(Ex: le décès d'un parent proche a-t-il été perçu comme un événement heureux ou malheureux à l'époque?; vous a-t-il affecté de 1 à 15 jours, de 15 jours à un mois ou plus de 1 mois?)

b) 17 questions attitudinales.

(Ex: Quelle est votre plus grande source de satisfaction dans la vie?: Le travail, la famille, les loisirs, autres.)

Problématique

Univers des impacts

Dans cet univers, les individus sont représentés par 27 variables; nous voulons donc répondre aux questions suivantes:

- A combien de dimensions indépendantes peut on réduire cet espace?
- Quelle est la signification des axes? Y a-t-il une correspondance physiologique qui supporte leur interprétation?
- En quoi les techniques d'AFC et d'ACP diffèrent-elles?

Les pourcentages de variance expliquée par les 4 premiers axes de l'Analyse en Composantes Principales sont respectivement de 53.00, 26.45, 6.60 et 3.50, soit un pourcentage cumulée de 89.54%.

La projection graphique des variables sur les axes 1 et 2 (Figure 1) fait ressortir une très forte corrélation entre le facteur 1 et l'ensemble des mesures de tension artérielle systolique et diastolique, tandis que les mesures de pouls sont caractéristiques de l'axe 2. Sur l'axe 3 s'opposent les mesures de tension systolique et diastolique, tandis que l'axe 4 permet de distinguer les mesures prises debout et couchées.

L'Analyse Factorielle des Correspondances pouvant être considérée de façon simpliste, comme une ACP appliquée à une matrice de distances Chi^2, nous avons transformé les données du référentiel en classes (5 classes pour la tension artérielle: <100, 100-110, 110-120, 120-130, >130, et 4 pour le pouls: <75, 75-80, 80-85, 85-90, >90).

Les pourcentages de variance expliquée par les 4 premiers axes sont respectivement de 13.97, 8.59, 7.13 et 5.34, soit un pourcentage cumulée de 35.03%. Il ne faut pas directement comparer les pourcentages expliqués par l'ACP et l'AFC, ni les interprétations, car dans un cas l'espace vectoriel est de 27 dimensions et de 99 dans l'autre; si le sujet est le même, les tableaux sont de factures différentes.

Sur le représentation graphique des axes 1 et 2 (Figure 2), nous retrouvons l'effet Guttman: le facteur 1 est caractérisé par les valeurs extrêmes de la tension artérielle, tandis que l'axe 2 oppose les extrêmes à la tendance centrale; le pouls est mal représenté sur ces axes. Le pouls est mieux représenté sur le facteur 3, tandis que sur le facteur 4, on retrouve la distinction entre les mesures prises en position debout et couchée.

Univers des attributs

Les questions que l'on se pose concernant ce référentiel:

- Peut-on filtrer les événements d'impact potentiel?
- Les valeurs test ont-elles un équivalent quand on passe de l'AFC-ACP aux analyses de variance traditionnelles?
- Les ANOVA contredisent-elles les Analyses Factorielles?
- La représentation AFC aide-t-elle à la "lecture" des Anova et en particulier des interactions?.

De façon générale, dans une analyse factorielle, les valeurs propres non nulles sont égales, que l'on parte de l'espace "variables" ou de l'espace "individus"; une simple transformation linéaire, associée à la matrice de départ, nous permet d'obtenir les directions propres cherchées dans l'autre espace. Il est possible alors de tester l'appartenance aux axes des classes formées par les variables illustratives de l'univers des impacts. Le test de T permet alors de filtrer les événements d'impact potentiel et de ne retenir que ceux contribuant "significativement" aux axes. En ACP, la projection graphique de ces classes nous permet de déterminer les tendances relationnelles avec les variables. Le fait d'avoir perdu un parent proche pendant l'année semble agir sur la tension artérielle, ainsi que le fait de trouver la plus grande source de satisfaction dans le travail par opposition à la source famille (Figure 3). La projection des axes 3 et 4 permet de préciser que la notion de décès influence la tension artérielle

diastolique. La projection barycentrique de l'univers filtré des conséquences sur les facteurs 1 et 2 de l'AFC nous permet de spécifier que la classe "décès" se positionne près des valeurs élevées (classe 5) de la tension artérielle par opposition aux autres classes, ainsi que la source de satisfaction "travail" par rapport aux autres sources (Figure 2).

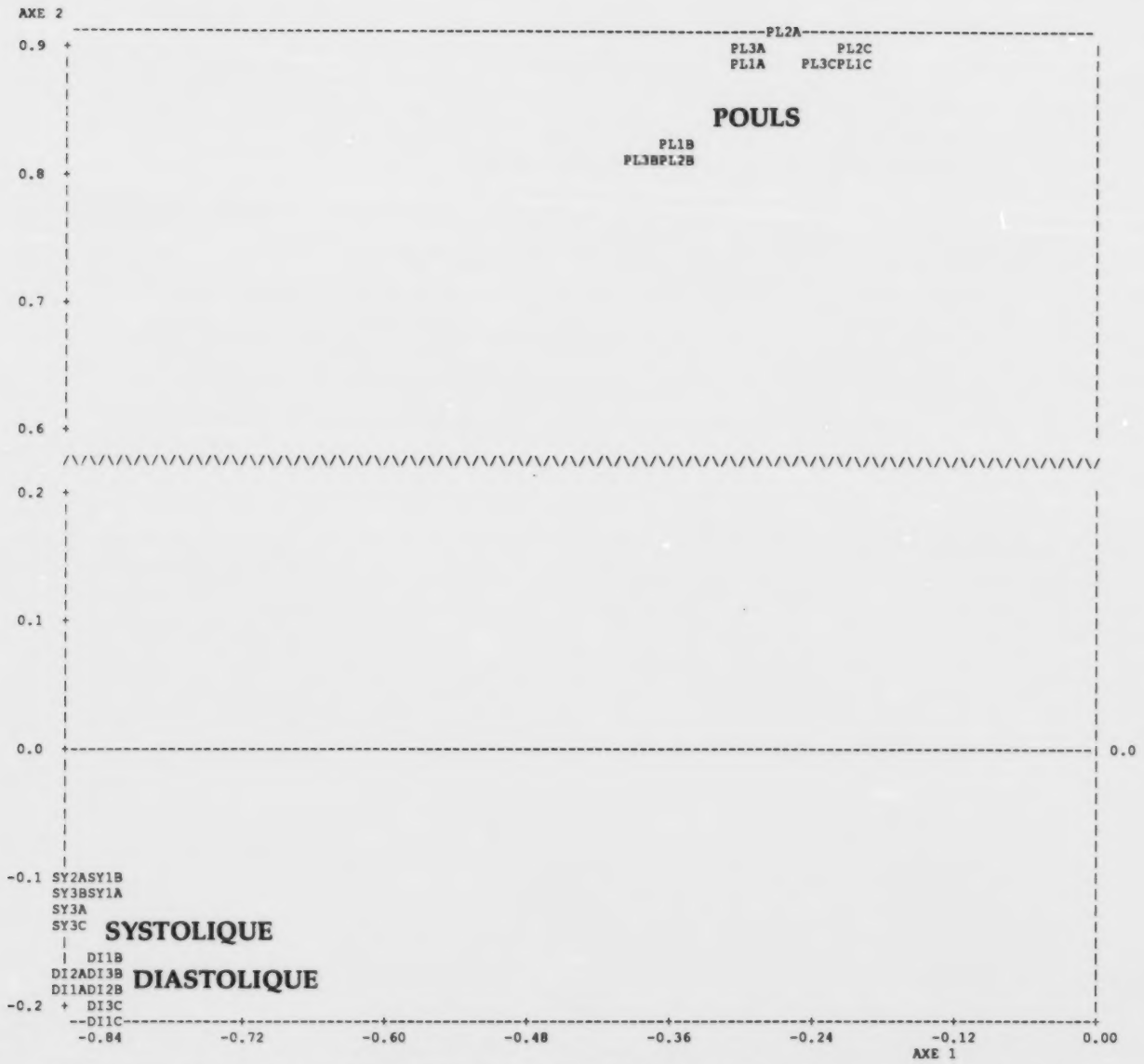
La représentation graphique de la tension artérielle systolique en fonction des sources de satisfaction et de la position de prise de tension (Figure 4) montre cette caractérisation de la source "travail". L'analyse de variance de la systolique (Figure 5) fait ressortir que la source de variation "satisfaction" est significative.

Conclusion

Nous nous garderons bien de conclure sur l'étude précitée, car on n'a pas tenu compte des variables confondantes tels que l'age, le sexe, l'indice d'obésité. Cet exemple n'avait pour but que d'illustrer notre propos sur l'utilisation complémentaire de l'analyse des données (ACP,AFC) et des analyses paramétriques en santé et sécurité du travail.

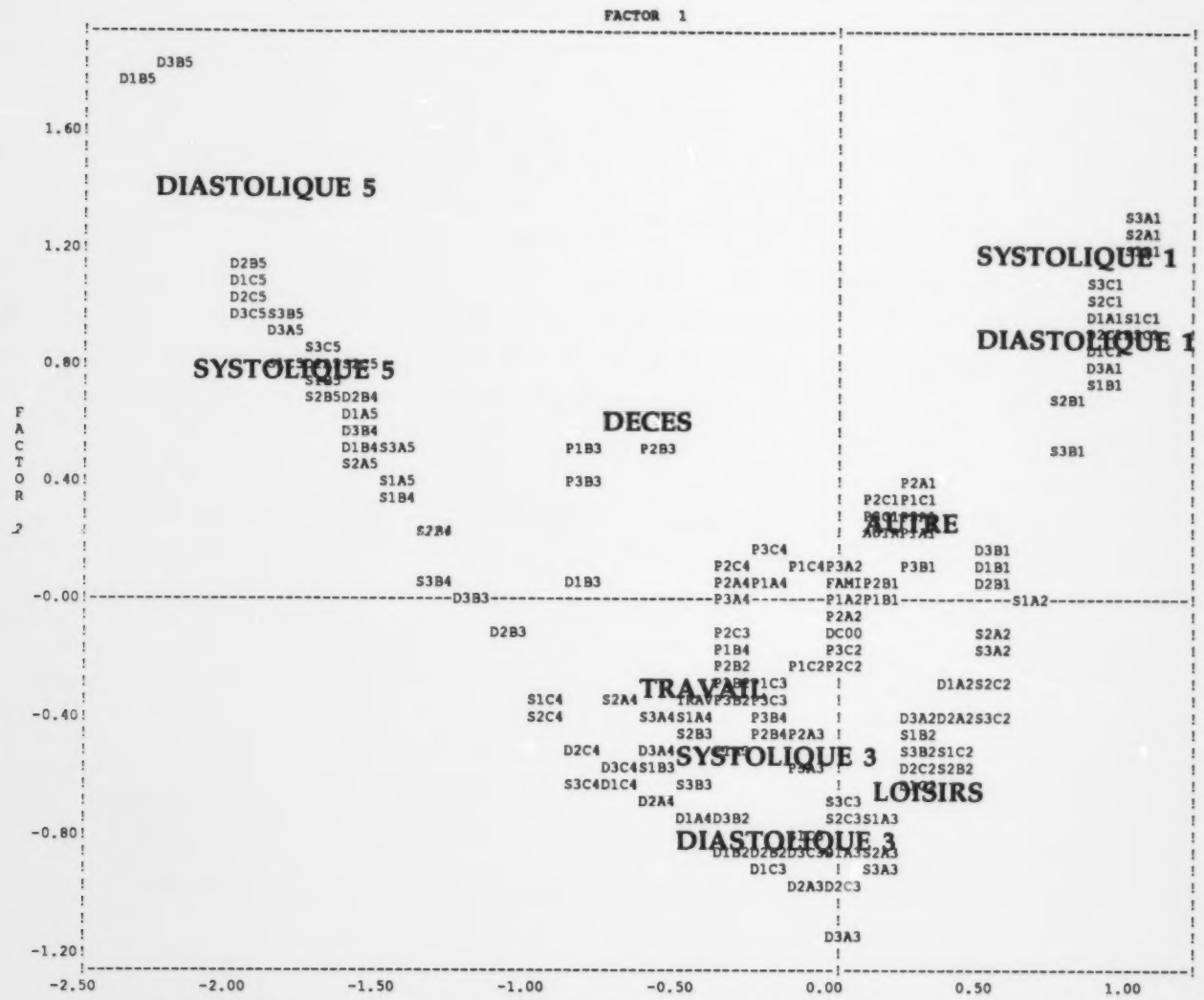
Lorsque nous disposons, dans une étude, de nombreuses données "secondaires", sur lesquelles il n'y a pas d'hypothèse à priori:

- L'analyse en composantes principales permet de réduire l'univers des conséquences et de filtrer l'univers des attributs, celles qui n'apportent que peu à l'étude; elle permet aussi de déterminer quelles variables de l'univers des impacts semblent impliquées dans le modèle.
- L'analyse Factorielle des correspondances permet d'émettre des hypothèses de relation entre les différents éléments d'une groupante et les variables d'impact.
- Les analyses paramétriques permettent d'infirmer ou confirmer les hypothèses énoncées, ainsi que les quantifier, quoique les énoncés probabilistes ne puissent être considérés exacts.



ANALYSE FACTORIELLE EN COMPOSANTES PRINCIPALES - EXAMEN PHYSIQUE
AXE 1 * AXE 2

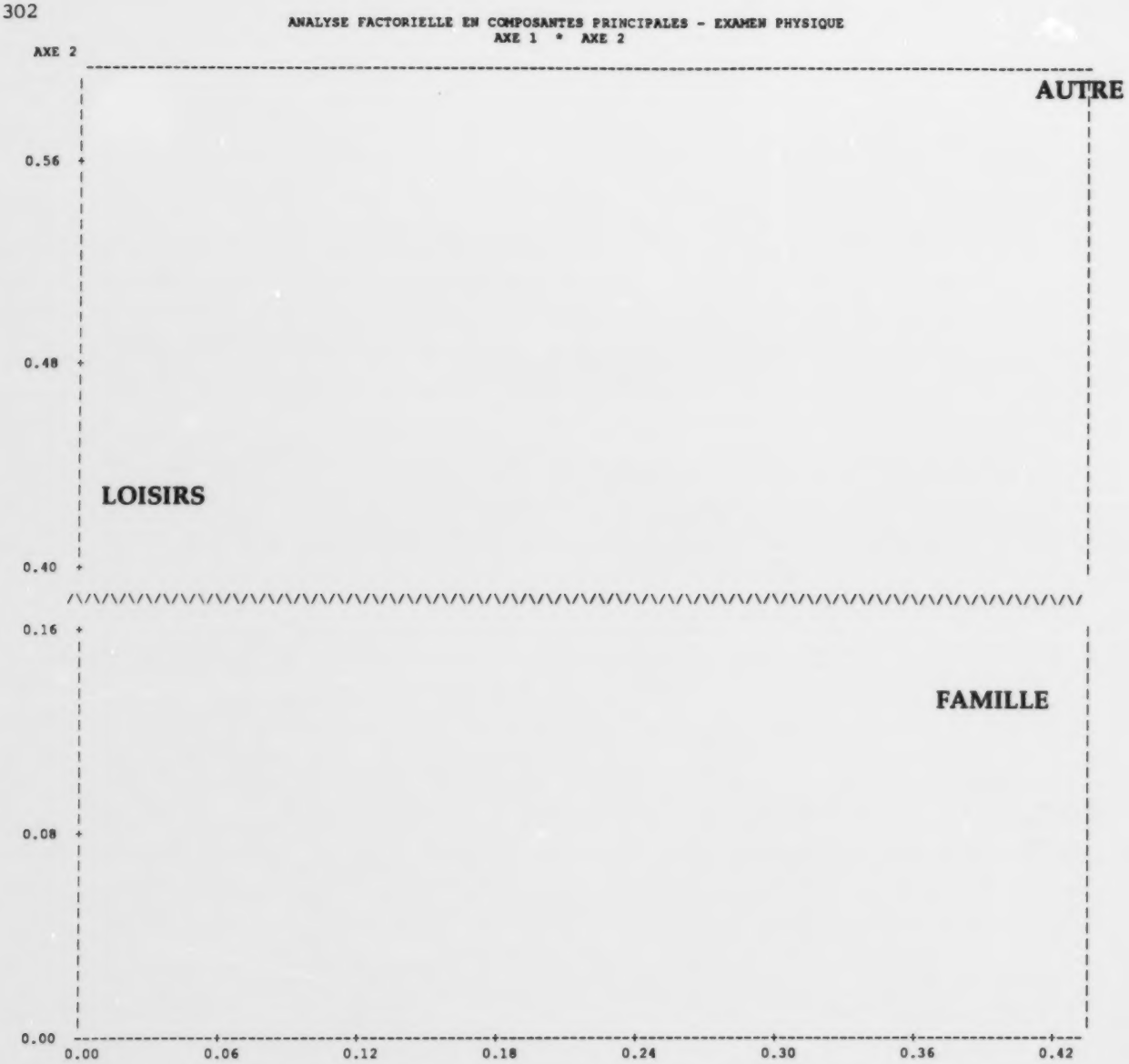
Figure 1: Les deux premiers facteurs de l'ACP illustrent la forte liaison des mesures de tension artérielle entre elles et avec l'axe 1; les mesures de pouls semblent peu corrélés aux mesures de tension, et fortement corrélés avec l'axe 2.



ANALYSE FACTORIELLE DE CORRESPONDANCES - EXAMEN PHYSIQUE

FACTOR 1 x FACTEUR 2

Figure 2: L'AFC permet la représentation simultanée de l'univers des impacts et de celui des attributs, ce qui en facilite l'interprétation. La projection barycentrique des individus éprouvés par le décès d'un parent proche dénote une tendance à une tension artérielle élévée chez ces individus.



ANALYSE FACTORIELLE EN COMPOSANTES PRINCIPALES - EXAMEN PHYSIQUE
AXE 1 * AXE 2

Figure 3: il est possible de positionner les tendances relationnelles des variables attributives entre elles sur les axes de l'ACP. L'interprétation doit tenir compte des orientations des variables d'impact projetées à la figure 1.

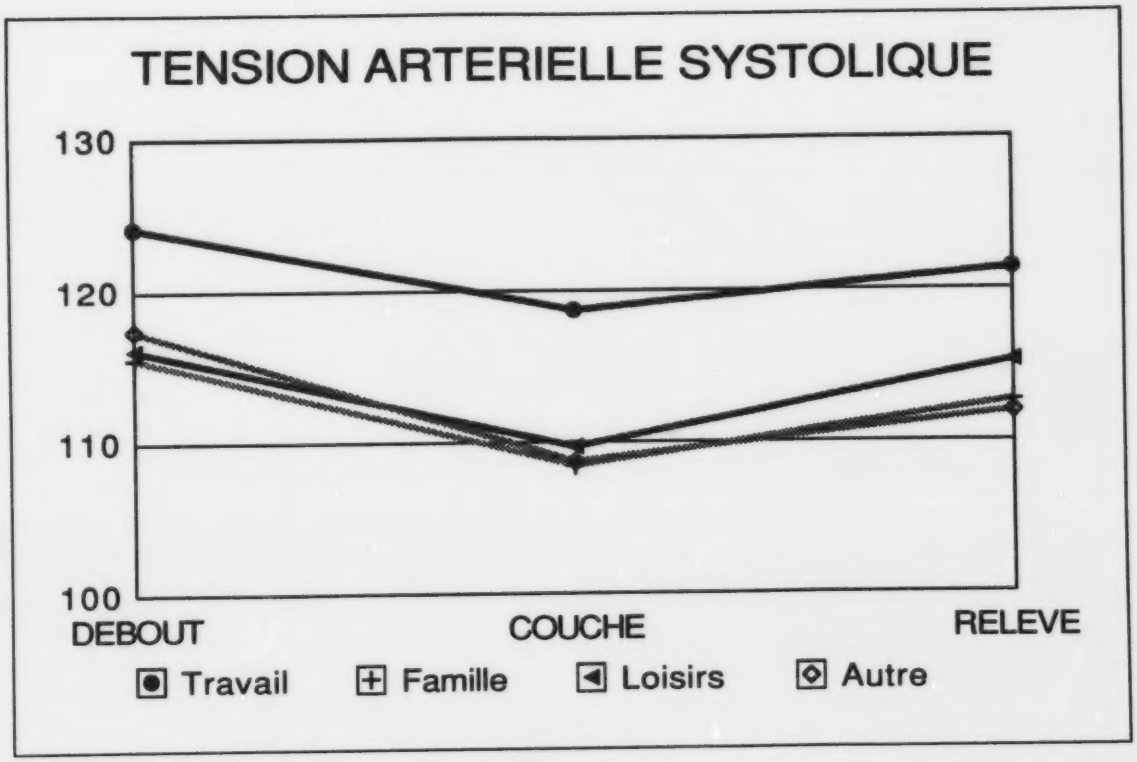


Figure 4
TENSION ARTÉRIELLE SYSTOLIQUE

SOURCE	SUM OF SQUARES	D.F.	MEAN SQUARE	F	TAIL PROB.
MEAN	9585461.18	1	9585461.18	6298.78	0.0000
SATISF	19525.88	3	6508.63	4.28	0.0061
ERROR 1	266314.58	175	1521.80		
POSITION	5860.15	2	2930.07	33.39	0.0000
PxS	403.20	6	67.20	0.77	0.5972
ERROR 2	30709.42	350	87.74		
MESURE	122.57	2	61.28	5.22	0.0059
MxS	60.96	6	10.16	0.87	0.5208
ERROR 3	4111.36	350	11.75		
PxM	393.83	4	98.46	8.76	0.0000
PxMxS	276.52	12	23.04	2.05	0.0182
ERROR 4	7866.30	700	11.23		

Figure 5

ÉTUDE DE LA VARIABILITÉ D'ÉCHANTILLONNAGE

Yves Béland[1]

La variance d'une estimation est une sorte de mesure de sa précision. Il est relativement peu coûteux de calculer des estimations de la variance pour des estimations simples lorsqu'on utilise des coefficients de pondération correspondant à l'inverse de la fraction de sondage et que l'on suppose un échantillonnage aléatoire simple. Toutefois, il est beaucoup plus onéreux d'estimer la variance, dans le cas d'estimations établies par la méthode itérative par le quotient, en prenant en considération le plan d'échantillonnage utilisé. Le but de cette étude était d'élaborer une méthode peu coûteuse de production de telles estimations de la variance. Pour ce faire, on a calculé des "facteurs d'ajustement".

1. INTRODUCTION

Les données qui sont recueillies à partir d'une enquête-échantillon sont toujours sujettes à certaines erreurs. On peut les diviser en deux types: les erreurs échantillonnales et les erreurs non-échantillonnales. L'erreur non-échantillonnale regroupe différentes erreurs comme les erreurs de couverture, les erreurs causées par les répondants (non-réponse, réponses biaisées) ainsi que les erreurs dues au traitement des données. L'erreur échantillonnale se divise en deux composantes: la variance et le biais. La variance mesure la variabilité de l'estimation des valeurs que prendrait une variable par rapport à sa valeur moyenne obtenue pour des répétitions hypothétiques du processus d'enquête. Le biais est la différence entre la valeur moyenne que prendrait une estimation dans l'hypothèse d'une répétition de la procédure d'enquête et la valeur réelle de la variable observée. L'erreur échantillonnale n'est pas connue. Cependant, la variance peut être estimée à l'aide des données recueillies dans l'enquête-échantillon. L'étude de la variabilité d'échantillonnage a été effectué afin d'estimer l'effet des méthodes d'échantillonnage et d'estimation sur les chiffres du recensement établis à partir des données-échantillon. Malheureusement, la variance ne donne aucune information sur le biais possible des données et sur l'erreur non-échantillonnale.

Durant le recensement de la population de 1986, 9 questions sur l'âge, le sexe, l'état matrimonial, la langue maternelle, le lien avec la personne repère du ménage, le type de logement et le mode d'occupation du logement ont été posées à toute la population. De plus, d'autres questions furent posées à un 1 logement privé occupé sur 5. Parmi celles-ci, on retrouvait les 9 mêmes questions et 23 autres questions socio-économiques. Ces données recueillies à partir de l'échantillon sont appelées données-échantillon 2B.

Afin d'attribuer des poids à chaque personne et à chaque ménage, l'ensemble du pays est divisé en régions de pondération (RP). Une région de pondération correspond au plus petit niveau géographique pour lequel on peut garantir qu'il y a concordance entre certaines estimations-échantillon et des chiffres connus relatif à la population. Elles sont formées de secteurs de

[1] Yves Béland, Division des méthodes d'enquêtes sociales
Statistique Canada, Ottawa

dénombrement contigus et on en compte 5341 au pays. A l'intérieur de ces régions, les personnes et les ménages qui font partie de l'échantillon sont classés dans des tableaux croisés dont les lignes et les colonnes sont définies à l'aide de caractéristiques connues pour l'ensemble de la population. Les poids des personnes et des ménages sont calculés avec la méthode itérative par le quotient. Les poids sont déterminés de façon à ce que le nombre estimé de personnes et de ménages qui se trouve dans chaque ligne ou chaque colonne des tableaux corresponde aux chiffres connus de la population.

Statistique Canada produit des milliers de tableaux à partir des données-échantillon du recensement de la population. Théoriquement, on pourrait fournir une mesure de précision (à savoir l'estimation de la variance d'échantillonnage) pour chaque estimation présentée dans ces tableaux. Cette mesure tiendrait compte à la fois du plan de sondage et de la méthode d'estimation. En pratique cependant, il est impossible d'établir de telles mesures pour toutes les estimations du recensement en raison du coût élevé du traitement des données. On calcule donc la variance d'échantillonnage pour un sous-ensemble seulement des estimations du recensement. Cela permet ensuite d'évaluer les effets combinés du plan de sondage et de la méthode d'estimation sur la variance d'échantillonnage. On peut ensuite ajuster les estimations simples de la variance d'échantillonnage, qui sont peu coûteuses à produire, pour tenir compte de cet effet et produire une estimation de la variance d'échantillonnage pour n'importe quelle estimation du recensement. Ce texte donne les résultats et décrit la méthodologie utilisée.

2. PRÉSENTATION DES RÉSULTATS

La racine carrée de la variance d'échantillonnage, soit l'erreur-type, peut être estimée à l'aide des données présentées dans les tableaux 1 et 2. Le tableau 1 donne des estimations non ajustées (simples) de l'erreur-type d'estimations-échantillon. Les valeurs présentées dans ce tableau ont été établies en supposant que la fraction de sondage était de 1/5 dans un échantillon aléatoire simple et qu'on avait utilisé une pondération simple de 5. Les erreurs-types présentées dans le tableau 1 sont fonction à la fois de l'importance du "total estimé" et de la population de la région. Par exemple, pour une estimation de 250 personnes dans une région comptant 1 000 personnes, l'erreur-type non ajustée est de 25. Le tableau 1 contient les erreurs-types associées à un nombre restreint de valeurs du "total estimé" et du nombre total de personnes, de ménages, de logements ou de familles dans la région. La formule ci-après peut être utilisée pour calculer l'erreur-type non ajustée de n'importe quel total estimé, pour toute région, quelle qu'en soit l'importance:

$$ETNA=\sqrt{\frac{4E\,(\,N-E\,)}{N}}$$

où ETNA est l'erreur-type non ajustée, E le total estimé et N le nombre total de personnes, de ménages, de logements, ou de familles dans la région. Par exemple, pour un total estimé de 750 personnes dans une région comptant 9 000 personnes, l'erreur-type non ajustée serait:

$$\sqrt{\frac{4(750)(9000-750)}{9000}} = 52$$

Le tableau 2 contient les facteurs d'ajustement (on appelle communément "effet du plan" le carré des facteurs d'ajustement) par lesquels on doit multiplier les erreurs-types non ajustées pour tenir compte des effets combinés du plan de sondage et de la méthode d'estimation. Pour calculer ces

facteurs d'ajustement, un échantillon de 401 RP a été tiré. L'échantillon a été réparti parmi les dix provinces de manière à obtenir de bonnes estimations de la variance d'échantillonnage au niveau provincial sans pour autant compromettre indûment la qualité des estimations au niveau national. Pour chaque RP dans l'échantillon, les estimations de la variance d'échantillonnage pour les estimations produites par la méthode itérative par le quotient ont été calculées pour différentes sous-catégories de toutes les caractéristiques énumérées dans le tableau 2. Les estimations de la variance d'échantillonnage aux niveaux provincial et national ont été obtenues en pondérant les estimations au niveau de la RP. Les facteurs d'ajustement pour chaque sous-catégorie de chacune des caractéristiques ont été calculés en divisant la racine carrée de ces estimations par l'erreur-type non ajustée correspondante. Des facteurs d'ajustement ont été calculés aux niveaux provincial et national pour chaque caractéristique en faisant la moyenne des facteurs d'ajustement de toutes les sous-catégories. Pour plus de renseignements sur la façon dont ces facteurs d'ajustement ont été calculés, consultez le chapitre 3 de ce texte.

Pour estimer l'erreur-type d'une estimation-échantillon donnée, l'utilisateur doit se reporter au tableau 2 pour obtenir le facteur d'ajustement s'appliquant à la caractéristique en question et multiplier ce facteur par l'erreur-type non ajustée correspondante dans le tableau 1. Lorsqu'une caractéristique ne figure pas au tableau 2, l'utilisateur doit se servir du facteur d'ajustement indiqué à la catégorie "Autres". Pour chaque caractéristique énumérée, le tableau 2 fournit les facteurs d'ajustement applicables aux niveaux national et provincial ainsi qu'au niveau de la RP. A moins que la région considérée ne soit plus petite que la province, il faut choisir la colonne "Facteur national ou provincial". Les facteurs d'ajustement pour les provinces sont fournis uniquement lorsqu'ils diffèrent de façon significative du facteur national. Cela s'est produit uniquement dans le cas des caractéristiques relatives à la langue maternelle. Si un facteur d'ajustement est requis pour une estimation se rapportant à une région plus petite qu'une province, les "centiles des facteurs au niveau de la RP" permettront d'obtenir une valeur plus précise. Les centiles donnent l'étendue de tous les facteurs d'ajustement au niveau des RP calculés dans l'étude pour les différentes sous-catégories d'une caractéristique. N% des facteurs d'ajustement au niveau de la RP se situaient en-dessous du $N^{ième}$ centile et 100 - N% se situaient au-dessus du $N^{ième}$ centile. Par exemple, 90% des facteurs d'ajustement au niveau de la RP étaient au-dessous du $90^{ième}$ centile et 10%, au-dessus. On décidera du centile à utiliser selon que l'on veut une estimation de l'erreur-type plus ou moins prudente. Par exemple, "100" nous donnera une estimation très prudente et "75", une estimation qu'il l'est moins.

Les règles suivantes doivent être appliquées pour calculer l'erreur-type ajustée :

(a) Lorsqu'on détermine l'erreur-type d'une estimation relative aux familles ou aux ménages, il faut utiliser le nombre de familles ou de ménages dans la région, pas le nombre de personnes, pour se reporter à la colonne qui convient dans le tableau 1.

(b) À moins que ce soit clairement spécifié dans le tableau 2, les caractéristiques de la famille de recensement et de la famille économique impliquant l'époux, l'épouse, le parent seul, ou la personne repère de la famille ont les mêmes facteurs d'ajustement que ceux mentionnés pour les caractéristiques de la population. Par exemple, le facteur d'ajustement de la caractéristique "plus haut niveau de scolarité atteint par l'époux, l'épouse, ou le parent seul de la famille de recensement" est le même que celui mentionné pour la caractéristique de la population "plus haut niveau de scolarité atteint".

(c) Pour le classement recoupé de deux ou plusieurs caractéristiques, il faut utiliser le facteur d'ajustement le plus élevé des caractéristiques en question.

(d) Tous les facteurs d'ajustement ont été calculés pour des estimations du nombre de personnes, de ménages, de logements, ou de familles, et non pour, par exemple, des unités comme les dollars. Par exemple, les facteurs d'ajustement pour la caractéristique "revenu du ménage" ont été calculés pour des estimations du nombre de ménages ayant un revenu compris dans différentes catégories telles 5 000 - 14 999 $, 35 000 - 44 999 $, etc., et non pour des estimations tel le revenu moyen du ménage.

L'exemple qui suit illustre la façon de calculer l'erreur-type ajustée. Supposons que la caractéristique étudiée est la population immigrante en Ontario. L'estimation du recensement de 1986 pour cette caractéristique était 2 081 200. Le chiffre de population obtenu pour l'Ontario au recensement de 1986 était 9 001 170. Comme aucun des deux chiffres n'est très proche des valeurs données dans le tableau 1, il faut appliquer la formule definie précédemment pour obtenir l'erreur-type non ajustée. Dans ce cas-ci, le résultat serait 2 530. Au tableau 2, on voit que le facteur d'ajustement au niveau provincial pour la caractéristique "immigrant" est 1.12. Il s'ensuit que l'erreur-type ajustée pour cette estimation est 2 530 X 1.12 = 2 834.

L'estimation-échantillon et l'erreur-type correspondante peuvent être utilisées pour construire un intervalle dans lequel on peut s'attendre, avec un degré de confiance donné, à trouver la valeur inconnue d'une caractéristique quelconque de la population. Pour cette étude, on a pris un échantillon parmi un grand nombre possible d'autres échantillons de même taille qu'on aurait pu sélectionner à l'aide du même plan de sondage. Les estimations obtenues à partir des données recueillies auprès des différents échantillons ne seraient pas les mêmes. Si l'on construisait des intervalles qui vont de deux erreurs-types au-dessous de l'estimation à deux erreurs-types au-dessus de cette dernière, alors environ 19 intervalles sur 20 comprendraient la "vraie" valeur qui aurait été obtenue au moyen d'un dénombrement complet. On appelle un tel intervalle un intervalle de confiance à 95% (19/20 = 95%). Toutefois, pour pouvoir garantir un niveau de confiance de 95%, les intervalles doivent être calculés en utilisant la "vraie" erreur-type des estimations-échantillon. Les erreurs-types ajustées, qui sont calculées à partir des tableaux 1 et 2, sont seulement des estimations des "vraies" erreurs-types. A noter cependant que les estimations-échantillon aux niveaux provincial et national devraient être assez proches des vraies erreurs-types pour permettre la construction, avec une précision raisonnable, d'intervalles de confiance à 95%. Au-dessous du niveau de la province, l'erreur-type ajustée ne serait peut-être pas assez précise pour permettre la construction d'intervalles semblables.

En prenant l'erreur-type calculée ci-dessus, on peut obtenir un intervalle de confiance à environ 95% pour le nombre d'immigrants en Ontario en procédant comme suit:
2 081 200 ± 2(2 834) soit 2 081 200 ± 5 668.

Tableau 1. Estimations non ajustées de l'erreur-type des estimations-échantillon

Total estimé	Nombre total de personnes, ménages, logements, où familles dans la région								
	500	1 000	2 500	5 000	10 000	25 000	50 000	100 000	250 000
50	15	15	15	15	15	15	15	15	15
100	18	19	20	20	20	20	20	20	20
250	22	25	30	30	30	30	30	30	30
500	0	30	40	40	45	45	45	45	45
1 000		0	50	55	60	60	65	65	65
2 500			0	70	85	95	95	100	100
5 000				0	100	130	130	140	140
10 000					0	150	180	190	200
25 000						0	220	270	300
50 000							0	320	400
100 000								0	490
250 000									0

Total estimé	Nombre total de personnes, ménages, logements, où familles dans la région					
	500 000	1 000 000	2 500 000	5 000 000	10 000 000	25 000 000
50	15	15	15	15	15	15
100	20	20	20	20	20	20
250	30	30	30	30	30	30
500	45	45	45	45	45	45
1 000	65	65	65	65	65	65
2 500	100	100	100	100	100	100
5 000	140	140	140	140	140	140
10 000	200	200	200	200	200	200
25 000	310	310	310	320	320	320
50 000	420	440	440	440	450	450
100 000	570	600	620	630	630	630
250 000	710	870	950	970	990	990
500 000	0	1 000	1 260	1 340	1 380	1 400
1 000 000		0	1 550	1 790	1 900	1 960
2 500 000			0	2 240	2 740	3 000
5 000 000				0	3 100	4 000
10 000 000					0	4 900

Tableau 2. Facteurs d'ajustement de l'erreur-type au niveau national ou provincial et centiles des facteurs au niveau de la région de pondération (RP)

Caractéristiques	Facteur national ou provincial	Centiles des facteurs au niveau de la RP					
		1	50	75	90	95	99	100
Caractéristiques de la population								
Âge								
. Groupes d'âge 0-4, 5-9, 10-14, 15-19, 20-24, 25-29	0.18	0.05	0.19	0.29	0.35	0.49	0.52	0.60
. Groupes d'âge 30-34, 35-44, 45-54, 55-59, 60-64, 5+, 15+	0.36	0.13	0.33	0.46	0.51	0.56	0.61	0.74
. Groupe d'âge 65+	0.00	-	-	-	-	-	-	-
Sexe	0.00	-	-	-	-	-	-	-
État matrimonial								
. Célibataire, marié (sauf séparé(e))	0.25	0.04	0.23	0.31	0.42	0.49	0.55	0.68
. Séparé(e), divorcé(e), veuf (veuve)	0.88	0.55	0.84	0.98	1.06	1.15	1.20	1.33
Plus haut niveau de scolarité atteint/plus haut grade, certificat, ou diplôme obtenu/nombre total d'années de scolarité	0.90	0.75	0.95	1.06	1.14	1.19	1.25	1.38
Principal domaine d'études	1.20	0.84	1.16	1.22	1.28	1.35	1.43	1.51
Mobilité								
. N'ayant pas déménagé	1.21	0.83	1.23	1.27	1.32	1.36	1.41	1.58
. Ayant déménagé (migrants,non-migrants)	1.61	0.90	1.60	1.75	1.85	1.97	2.09	2.21
Période d'immigration								
. Avant 1946, 1946-1966	0.98	0.76	1.02	1.10	1.22	1.37	1.45	1.62
. 1967-1977, 1978-1982, 1983-1986	1.51	0.80	1.45	1.55	1.78	1.90	2.11	2.20
Âge à l'immigration	1.10	0.71	1.15	1.29	1.38	1.44	1.54	1.67
Lieu de naissance								
. Au Canada	1.09	0.82	1.08	1.16	1.18	1.20	1.21	1.33
. À l'étranger	1.35	1.11	1.34	1.43	1.60	1.67	1.75	1.91
Population immigrante/non immigrante	1.12	0.81	1.10	1.24	1.38	1.46	1.52	1.69
Citoyenneté								
. Du Canada, par naissance	1.13	0.88	1.14	1.17	1.20	1.27	1.32	1.58
. D'un autre pays	1.59	1.04	1.40	1.65	1.88	1.95	2.12	2.30

Tableau 2. Facteurs d'ajustement de l'erreur-type au niveau national ou provincial et centiles des facteurs au niveau de la région de pondération (RP) - suite

Caractéristiques	Facteur national ou provincial	Centiles des facteurs au niveau de la RP						
		1	50	75	90	95	99	100
Origine ethnique								
. Français, anglais	1.20	0.73	1.16	1.25	1.31	1.40	1.46	1.65
. Autres	1.65	1.07	1.57	1.70	1.89	1.99	2.11	2.45
Langue parlée à la maison								
. Français, anglais, français et anglais, anglais et autres	1.12	0.50	1.09	1.35	1.75	1.89	2.09	2.19
. Autres	1.76	0.99	1.68	1.89	2.01	2.20	2.41	2.66
Langue officielle								
. Français, anglais, français et anglais	1.05	0.69	1.01	1.18	1.31	1.42	1.58	1.75
. Autres	1.49	0.90	1.50	1.68	1.76	1.79	1.91	2.01
Langue maternelle - anglais								
. Terre-Neuve, Ile-du-Prince-Édouard, Nouvelle-Écosse, Colombie-Britannique	0.92	0.24	0.96	1.45	1.62	1.90	2.23	2.45
. Québec	1.15	0.18	1.10	1.51	1.76	1.81	1.99	2.21
. Autres provinces	0.45	0.12	0.48	0.71	0.96	1.12	1.38	1.68
. Canada	0.53	-	-	-	-	-	-	-
Langue maternelle - français								
. Québec	0.42	0.14	0.45	0.52	0.61	0.76	0.91	1.19
. Nouveau-Brunswick	0.75	0.19	0.79	0.98	1.24	1.60	1.84	2.04
. Autres provinces	1.04	0.09	1.12	1.49	1.71	1.89	2.06	2.40
. Canada	0.77	-	-	-	-	-	-	-
Langue maternelle - autre	1.70	0.73	1.63	2.11	2.44	2.51	2.60	2.70
Industrie/profession	0.92	0.25	0.80	1.13	1.25	1.31	1.38	1.67
Travail en 1985	0.89	0.62	0.92	1.14	1.22	1.29	1.31	1.45
Semaines travaillées en 1985	0.94	0.68	0.99	1.18	1.29	1.33	1.39	1.69
Heures travaillées durant la semaine de référence	0.83	0.63	0.85	1.01	1.14	1.19	1.24	1.51
Date du dernier travail								
. En 1986, en 1985, avant 1985	0.89	0.60	0.94	0.99	1.05	1.11	1.20	1.33
. N'a jamais travaillé	1.18	0.80	1.15	1.34	1.43	1.50	1.67	1.82

Tableau 2. Facteurs d'ajustement de l'erreur-type au niveau national ou provincial et centiles des facteurs au niveau de la région de pondération (RP) - suite

Caractéristiques	Facteur national ou provincial	Centiles des facteurs au niveau de la RP						
		1	50	75	90	95	99	100
Catégorie de travailleur								
. Travailleur rémunéré	0.72	0.56	0.75	0.86	0.93	0.95	0.98	1.09
. Travailleur autonome (entreprise ou ferme non constituée en société, travailleur familial non rémunéré)	0.93	0.68	0.96	1.08	1.13	1.15	1.18	1.31
Activité								
. Occupé	0.75	0.59	0.76	0.83	0.86	0.91	0.93	1.04
. En chômage	1.06	0.76	1.04	1.14	1.20	1.27	1.38	1.53
. Inactif	1.25	0.91	1.30	1.43	1.50	1.58	1.63	1.84
Principal source de revenu								
. Salaires et traitements	0.65	0.42	0.67	0.80	0.85	0.87	0.92	0.99
. Autres	1.05	0.71	1.00	1.12	1.17	1.20	1.24	1.48
Incapacité								
. Limité à la maison, à l'école, au travail	0.94	0.69	0.96	1.11	1.29	1.34	1.42	1.69
. Pas limité	0.61	0.41	0.58	0.69	0.74	0.78	0.81	0.84
Situation dans la famille de recensement								
. Époux, épouse, enfant	0.20	0.05	0.20	0.24	0.26	0.28	0.31	0.34
. Parent seul de sexe féminin	0.45	0.14	0.43	0.51	0.55	0.61	0.68	0.81
. Parent seul de sexe masculin, personne hors famille	0.68	0.35	0.65	0.79	0.89	0.99	1.14	1.32
Situation dans la famille économique								
. Époux, épouse	0.14	0.06	0.16	0.21	0.28	0.34	0.36	0.42
. Parent seul, enfant	0.32	0.16	0.34	0.39	0.44	0.47	0.53	0.68
. Autres membres de la famille	0.74	0.24	0.70	0.84	1.03	1.09	1.18	1.31
Nombre de personnes dans la famille de recensement	0.04	0.00	0.00	0.05	0.07	0.09	0.11	0.13
Nombre de personnes dans la famille économique	0.18	0.08	0.19	0.24	0.33	0.41	0.45	0.71
Âge de l'époux, de l'épouse, ou de la personne repère de la famille économique	1.42	0.80	1.37	1.53	1.60	1.78	1.91	2.08
Autres caractéristiques de le population	1.00	-	-	-	-	-	-	-

Tableau 2. Facteurs d'ajustement de l'erreur-type au niveau national ou provincial et centiles des facteurs au niveau de la région de pondération (RP) - suite

Caractéristiques	Facteur national ou provincial	Centiles des facteurs au niveau de la RP						
		1	50	75	90	95	99	100
Caractéristiques du ménage et du logement								
Type de construction								
. Maisons individuelles	0.33	0.05	0.35	0.55	0.67	0.75	0.89	1.08
. App. dans un immeuble < 5 étages	0.57	0.12	0.56	0.70	0.83	0.99	1.26	1.44
. Autres	0.91	0.18	0.88	0.99	1.18	1.23	1.32	1.51
Mode d'occupation	0.00	-	-	-	-	-	-	-
Période de construction	0.78	0.61	0.75	0.82	0.89	0.99	1.24	1.49
Système de chauffage principal/ principal combustible pour le chauffage	0.87	0.18	0.86	1.04	1.12	1.25	1.32	1.47
Système de chauffage central								
. Avec chauffage central	0.42	0.09	0.38	0.54	0.60	0.70	0.89	1.19
. Sans chauffage central	0.78	0.23	0.79	0.91	1.03	1.12	1.20	1.39
Taille du ménage								
. Une personne	0.00	-	-	-	-	-	-	-
. Autres	0.76	0.19	0.72	1.09	1.17	1.21	1.30	1.53
Nombre de pièces	0.80	0.57	0.78	0.90	0.97	1.10	1.20	1.44
Âge du soutien de ménage								
. 25-34, 55-64, 65-74, 75+	0.25	0.06	0.24	0.35	0.48	0.53	0.62	0.94
. 0-24, 35-44, 45-54	0.92	0.38	0.90	1.05	1.14	1.21	1.30	1.49
Sexe du soutien de ménage								
. Homme	0.20	0.09	0.24	0.31	0.34	0.36	0.37	0.42
. Femme	0.47	0.16	0.43	0.54	0.64	0.74	0.89	1.09
Loyer brut/pourcentage du loyer brut dans le revenu du ménage	0.75	0.48	0.79	0.91	0.94	0.96	1.01	1.21
Principales dépenses de propriété/pourcentage des principales dépenses de propriété dans le revenu du ménage	0.84	0.62	0.87	0.95	1.01	1.04	1.11	1.29
Revenu du ménage	0.75	0.51	0.73	0.82	0.90	0.95	1.03	1.17
Valeur du logement	0.90	0.67	0.91	1.00	1.05	1.12	1.18	1.32
Copropriété enregistrée								
. Enregistrée	0.63	0.18	0.59	0.84	0.93	1.11	1.30	1.48
. Non enregistrée	0.15	0.07	0.14	0.19	0.28	0.39	0.47	0.59

Tableau 2. Facteurs d'ajustement de l'erreur-type au niveau national ou provincial et centiles des facteurs au niveau de la région de pondération (RP) - suite

Caractéristiques	Facteur national ou provincial	Centiles des facteurs au niveau de la RP						
		1	50	75	90	95	99	100
Genre de ménage - ménages unifamiliaux								
. Sans "autres personnes"	0.22	0.05	0.20	0.27	0.33	0.36	0.40	0.56
. Avec "autres personnes"	0.50	0.20	0.48	0.61	0.72	0.74	0.79	0.90
Genre de ménage - ménages non familiaux	0.00	-	-	-	-	-	-	-
Genre de ménage - autres	1.12	0.54	1.05	1.26	1.40	1.51	1.67	1.91
Autres caractéristiques des ménages et des logements	1.00	-	-	-	-	-	-	-
Caractéristiques de la famille de recensement								
Structure de la famille								
. Époux-épouse	0.20	0.09	0.21	0.26	0.29	0.33	0.36	0.42
. Parent seul de sexe masculin	0.64	0.21	0.62	0.81	0.84	0.91	1.04	1.25
. Parent seul de sexe féminin	0.46	0.19	0.45	0.57	0.65	0.69	0.74	0.91
Genre de famille								
. Famille principale	0.23	0.04	0.24	0.28	0.31	0.34	0.39	0.52
. Famille secondaire	0.90	0.62	0.93	1.15	1.28	1.33	1.40	1.49
Groupes d'âge des enfants à la maison	0.78	0.40	0.70	0.91	0.98	1.09	1.19	1.45
Activité de l'époux, de l'épouse, ou du parent seul								
. Époux, parent seul, époux-épouse actifs	0.40	0.23	0.43	0.50	0.55	0.59	0.71	0.93
. Epouse active	0.61	0.41	0.60	0.68	0.74	0.78	0.82	1.15
. Autres	0.72	0.30	0.68	0.80	0.90	0.99	1.12	1.38
Travail en 1985 de l'époux, de l'épouse, ou du parent seul								
. À travaillé en 1985	0.48	0.11	0.45	0.50	0.54	0.57	0.59	0.63
. N'a pas travaillé en 1985	0.93	0.60	0.90	1.04	1.18	1.26	1.30	1.43
Autres caractéristiques de la famille de recensement	1.00	-	-	-	-	-	-	-

Tableau 2. Facteurs d'ajustement de l'erreur-type au niveau national ou provincial et centiles des facteurs au niveau de la région de pondération (RP) - fin

Caractéristiques	Facteur national ou provincial	Centiles des facteurs au niveau de la RP						
		1	50	75	90	95	99	100
Caractéristiques de la famille économique								
Structure de la famille économique								
. Familles époux-épouse	0.29	0.13	0.30	0.36	0.48	0.56	0.68	0.91
. Autres	0.56	0.35	0.50	0.66	0.81	0.90	1.06	1.28
Langue maternelle de la personne repère de la famille: Anglais								
. Terre Neuve, Ile-du-Prince-Édouard, Colombie-Britannique	0.25	0.09	0.20	0.31	0.45	0.66	0.91	1.43
. Québec	0.49	0.25	0.47	0.50	0.69	0.83	1.05	1.53
. Autres provinces	0.18	0.07	0.19	0.22	0.24	0.27	0.31	0.49
. Canada	0.27	-	-	-	-	-	-	-
Langue maternelle de la personne repère de la famille: Français								
. Québec	0.12	0.04	0.13	0.17	0.21	0.29	0.36	0.51
. Autres provinces	0.88	0.30	0.90	1.07	1.21	1.28	1.35	1.69
. Canada	0.40	-	-	-	-	-	-	-
Langue maternelle de la personne repère de la famille: Ni le français, ni l'anglais								
. Terre-Neuve, Nouvelle-Écosse	0.75	0.38	0.74	0.80	0.91	0.99	1.10	1.38
. Autres provinces	0.50	0.21	0.45	0.57	0.82	0.84	0.99	1.47
. Canada	0.56	-	-	-	-	-	-	-
Autres caractéristiques de la famille économique	1.00	-	-	-	-	-	-	-

3. MÉTHODOLOGIE DE L'ÉTUDE DE LA VARIABILITÉ D'ÉCHANTILLONNAGE

3.1 INTRODUCTION

En se basant sur les données-échantillon du recensement de la population de 1986, Statistique Canada a produit des milliers d'estimations. Avoir eu un budget illimité en ressources informatiques, les estimations de la variance auraient été calculées pour toutes les estimations-échantillon à l'intérieur de chacune des RP du pays. Etant donné les contraintes de coût, les estimations de la variance ont été calculées pour un échantillon d'estimations-échantillon et un échantillon de RP.

3.2 STRATIFICATION DES RP

Tout d'abord, pour les besoins de l'étude, les RP ayant aucune population à l'intérieur ont été exclues. A l'intérieur de ces RP, seuls les logements privés ont été considérés.

La population de RP a tout d'abord été divisée en groupes de divisions de recensement contigues. Une division de recensement englobe entre 15 et 25 RP et on compte 266 divisions de recensement au Canada. Le pays a été subdivisé en 26 groupes de divisions de recensement et chaque groupe respectait les limites provinciales. Pour les besoins de l'étude, le Yukon et les Territoires du Nord-Ouest ont été inclus dans la Colombie-Britannique. Les groupes ont été formés en considérant la position géographique des divisions de recensement. Par exemple, la région de Montréal forme un groupe de même que le nord de l'Ontario. Par la suite, les RP à l'intérieur de chaque groupe (sauf pour les groupes de Montréal, Toronto et Vancouver) ont été classées selon un indice rural/urbain. Une RP a été considérée rurale si au moins 33% des secteurs de dénombrement à l'intérieur de cette RP étaient ruraux. Donc, la population de RP a été stratifiée en 49 strates.

Dans le but d'améliorer l'efficacité de la stratification, certaines RP ont été classées dans des strates autoreprésentatives. En fait, excepté pour l'Île-du-Prince-Édouard, une strate autoreprésentative a été créée dans chaque province. Au total, 39 RP ont été classées dans ces strates. Les critères qui ont été utilisés pour identifier les RP inclues dans ces strates sont décrits dans la section 3.4 de ce texte.

3.3 RÉPARTITION DE L'ÉCHANTILLON DE RP

Dans le but d'effectuer une répartition plus adéquate de l'échantillon de RP, une étude pilote a été effectuée. Le but de cette étude pilote était d'estimer la variance pour quelques caractéristiques à l'intérieur de chaque strate. A l'exception des strates de Montréal et de Toronto et des strates autoreprésentatives, un échantillon aléatoire simple sans remise de 4 RP a été sélectionné dans chacune des 47 strates. 8 RP ont été sélectionnées dans Montréal et Toronto. Les 39 RP inclues dans les strates autoreprésentatives ont demeuré dans les 49 strates standard pour les besoins de l'étude pilote. L'échantillon de l'étude pilote était finalement composé de 204 RP.

Dans chaque strate, les estimations de la variance des estimations-échantillon ont été calculées pour quelques 460 caractéristiques. Quelques-unes de ces estimations ont servi à répartir 4 tailles possibles d'échantillon de RP (350, 400, 450 et 500) parmi les 49 strates. Les strates autoreprésentatives ont été échantillonnées à 100%.

Etant donné que les tailles des strates variaient quelque peu entre elles, les méthodes standard pour répartir un échantillon parmi les strates étaient inadéquates. En effet, une répartition selon

Neyman a tendance à minimiser les coefficients de variation (CV) au niveau national mais peut, d'un autre côté, fournir d'imposants CV pour les petites strates. Par contre, une répartition qui a tendance à fournir des CV égaux au niveau des strates peut donner des grands CV au niveau national. Pour cette étude, un compromis entre ces deux méthodes de répartition a été effectué. Nous avons utilisé une méthode appelée "la répartition exponentielle" (voir Bankier [2]). La répartition a été effectuée en supposant que la variance des estimations de la variance dépendait seulement de l'échantillonnage des RP. La variance introduite par l'échantillonnage des ménages (1 sur 5) à l'intérieur des secteurs de dénombrement au niveau d'une RP n'a pas été considérée. Estimer la variance de l'estimation de la variance au niveau d'une RP d'une estimation obtenue par la méthode itérative par le quotient aurait été très difficile.

En utilisant les estimations de la variance d'un échantillon aléatoire simple de 30 caractéristiques calculées dans l'étude pilote, la méthode de la répartition exponentielle a été appliquée, à différentes puissances ($0 \leq q \leq 1$), à chacune de ces caractéristiques. Ces caractéristiques ont été sélectionnées parmi les caractéristiques des personnes et des ménages.

Tout d'abord, il a été décidé qu'un échantillon de 400 RP pour l'étude procurerait suffisamment de précision au niveau des strates considérant le budget disponible en ressources informatiques. Dans le but de respecter la taille totale de 400 RP pour l'échantillon, les RP qui ont été classées dans les strates autoreprésentatives ont été retirées des 400 RP. Il a également été décidé d'utiliser la puissance $q=0.5$ pour répartir l'échantillon de 361 RP parmi les strates. Les tailles finales des strates ont été obtenues en effectuant la moyenne sur les 30 possibilités (associées à chaque caractéristique échantillonnée). Finalement, la taille finale de l'échantillon était de 401 RP (362 + 39) et les RP de l'étude pilote ont été inclues dans l'échantillon final.

Pour différentes puissances de la répartition, la figure 1 trace le graphique des moyennes aux niveaux provincial et national des CV des estimations de la variance des 30 caractéristiques échantillonnées. Il montre qu'avec la répartition selon Neyman ($q=1$), les CV pour les petites provinces sont passablement élevés comparativement à ceux obtenus avec une répartition avec CV égaux pour les strates ($q=0$).

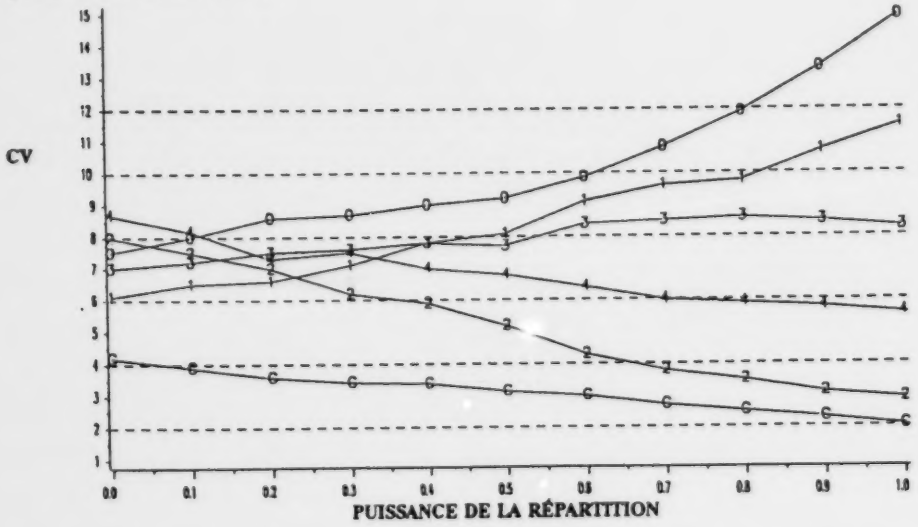


FIGURE 1. CV aux Niveaux Provincial et National pour Différentes Puissances: 0, Île-du-Prince-Édouard; 1, Nouveau-Brunswick; 2, Ontario; 3, Manitoba; 4, Colombie-Britannique; C, Canada.

3.4 ESTIMATION DE LA VARIANCE

La procédure d'estimation de la variance d'échantillonnage au niveau d'une RP qui a été utilisée pour cette étude est basée sur le travail de M. Bankier (voir Bankier [3]). Elle correspond à l'estimation de la variance après p itérations de la méthode itérative par le quotient obtenue après avoir linéarisé les données en utilisant les séries de Taylor. Dans ce cas-ci, p est défini comme étant le nombre d'itérations qu'il a fallu effectuer pour déterminer les poids finaux des personnes et des ménages avec la méthode itérative par le quotient. Cette procédure d'estimation prend en considération la stratification d'une RP en secteurs de dénombrement et elle prend également en considération l'échantillonnage de grappes de personnes à l'intérieur des secteurs de dénombrement.

Pour chaque RP sélectionnée dans l'étude pilote, les estimations de la variance ont été calculées en utilisant cette approche pour quelques 460 caractéristiques. Pour l'autre partie de l'échantillon de RP, les estimations de la variance ont été calculées pour un échantillon de caractéristiques. Cet échantillon sera décrit un peu plus loin dans ce texte.

Au niveau provincial, l'estimation de la variance Horvitz-Thompson (H-T) a été utilisée comme information auxiliaire dans un estimateur combiné par le quotient. Ces variances H-T ont été calculées au niveau du secteur de dénombrement à partir de la base de données des données-échantillon et elles ont été cumulées par la suite au niveau des RP et au niveau provincial. Les poids utilisés au niveau du secteur de dénombrement pour l'estimation de la variance H-T correspondent à l'inverse de la fraction de sondage finale des ménages.

Pour une certaine caractéristique d'intérêt, pour la $i^{ième}$ RP de la $h^{ième}$ strate de la province p, posons $\hat{y}_{phi}{}^{r}$ comme étant l'estimation obtenue avec la méthode itérative par le quotient et $\hat{y}_{phi}$ l'estimation H-T (en supposant un plan simple et une méthode d'estimation simple). Posons $\hat{v}_{phi}$ comme étant l'estimation quasiment non-biaisée au niveau de la RP de la variance de $\hat{y}_{phi}{}^{r}$ et t_{phi} l'estimation non-biaisée de la variance de $\hat{y}_{phi}$. Posons également H_p le nombre de strates autoreprésentatives dans la $p^{ème}$ province, N_{ph} le nombre de RP dans la population dans la $h^{ième}$ strate de la province p et n_{ph} le nombre de RP dans l'échantillon. L'estimation de la variance au niveau provincial a été calculée de la manière suivante:

$$\hat{V}_p^{c} = T_p \left(\frac{\hat{V}_p}{\hat{T}_p} \right)$$

$$\text{où} \quad T_p = \sum_{h}^{H_p} \sum_{i}^{N_{ph}} t_{phi} \ , \quad \hat{T}_p = \sum_{h}^{H_p} \left(\frac{N_{ph}}{n_{ph}} \right) t_{ph} \ , \quad t_{ph} = \sum_{i}^{n_{ph}} t_{phi} \ ,$$

$$t_{phi} = \sum_{j=1}^{K_{phi}} t_{phij} \ , \quad t_{phij} = \frac{M_{phij}^2}{m_{phij}} \left(1 - f_{phij} \right) s_{phij}^2 \ , \quad f_{phij} = \frac{m_{phij}}{M_{phij}} \ ,$$

$$\hat{V}_p = \sum_h^{H_p} \left(\frac{N_{ph}}{n_{ph}} \right) \sum_i^{n_{ph}} \hat{V}_{phi} \ ,$$

où M_{phij} est le nombre de ménages dans la population dans le $j^{ième}$ secteur de dénombrement de la $i^{ième}$ RP de la $h^{ième}$ strate de la province p et m_{phij} est le nombre de ménages dans l'échantillon de 1 ménage sur 5. Egalement, K_{phi} est le nombre de secteurs de dénombrement dans une RP.

Dans le but d'isoler les RP qui pourraient fournir des estimations de la variance trop grandes pour certaines caractéristiques, une strate autoreprésentative a été créée dans chaque province (excepté pour l'IPE). Ces strates autoreprésentatives incluaient les RP pour lesquelles la sommation des estimations de la variance H-T sur les 460 caractéristiques était très élevée. Ces RP ont été déterminées de la manière suivante. Posons Q3-Q1 comme étant l'écart interquartile de toutes les sommations des estimations de la variance H-T sur toutes les caractéristiques pour chaque RP (écart entre les $25^{ième}$ et $75^{ième}$ centiles). A l'intérieur de chaque province, excepté pour l'Ontario et le Québec, une RP était inclue dans la strate autoreprésentative si la sommation des estimations de la variance H-T pour cette RP était supérieure à Q3+1.5*(écart interquartile). Pour les provinces de l'Ontario et du Québec, une RP était inclue dans la strate autoreprésentative si la sommation était supérieure à Q3+3*(écart interquartile). Finalement, 39 RP ont été classées dans des strates autoreprésentatives.

Les estimations de la variance qui ont été calculées pour chaque RP de l'étude pilote (460 caractéristiques) ont servi à vérifier si les groupes de caractéristiques qui avaient été formés dans l'étude de la variabilité d'échantillonnage de 1981 pouvaient être utilisés lors de l'étude de 1986. En fait, les effets du plan de sondage ont été estimés pour chaque caractéristique de chaque RP sélectionnée dans l'étude pilote et les effets du plan nationaux et provinciaux ont été calculés pour chacun des groupes de caractéristiques de 1981. La plupart des groupes formés pour l'étude de 1981 ont été retenus pour l'étude de 1986, seulement quelques groupes ont été divisés en 2 ou plusieurs sous-groupes. Ces groupes qui ont été divisés sont ceux qui avaient une grande variabilité entre les effets du plan d'un même groupe.

Les effets du plan nationaux ont été estimés pour chaque groupe de caractéristiques en prenant la moyenne des effets du plan individuels. Cependant, certains effets du plan nationaux de quelques groupes ont été séparés en effets du plan provinciaux. Cette situation s'est produite pour les groupes où il y avait une grande variabilité entre les effets du plan provinciaux. Pour une certaine caractéristique, posons H_{TA} comme étant le nombre de strates autoreprésentatives (une pour chaque province excepté pour l'IPE). Les effets du plan nationaux (d_{nat}), provinciaux (d_p) et au niveau d'une RP (d_{phi}) ont été estimés par

$$d_{nat} = \frac{\sum_p \hat{V}_p^c + \sum_p^{H_{TA}} \hat{V}_p^{TA}}{t_{nat}^*} \ ,$$

$$\partial_p = \frac{\hat{V}_p^C + \hat{V}_p^{TA}}{t_p^*} \quad , \quad \partial_{phi} = \frac{\hat{v}_{phi}}{t_{phi}^*} \quad ,$$

où $\hat{V}_p^{TA}$ est la somme des estimations de la variance de l'estimation obtenue avec la méthode itérative par le quotient de toutes les RP de la $p^{ième}$ strate autoreprésentative (province). Les variances t_{nat}^*, t_p^* et t_{phi}^* sont les variances H-T aux niveaux national, provincial et de la RP qui ont été calculées avec la formule ETNA donnée à la section 2. Ces variances H-T sont basées sur la valeur du total estimé obtenue avec la méthode itérative par le quotient en utilisant l'échantillon de 401 RP (et non la population totale du pays).

A chaque facteur d'ajustement national ou provincial associé aux groupes de caractéristiques fournis dans le tableau 2, des centiles sont donnés. Ces centiles ont été pondérés au niveau d'une RP par l'inverse de la fraction de sondage de la RP. Ces centiles donnent une idée de l'étendue des effets du plan au niveau des RP de toutes les caractéristiques à l'intérieur d'un groupe. N% des facteurs d'ajustement au niveau de la RP se situaient en-dessous du $N^{ième}$ centile et 100 - N% se situaient au-dessus du $N^{ième}$ centile. Par exemple, 90% des facteurs d'ajustement au niveau de la RP étaient au-dessous du $90^{ième}$ centile et 10%, au-dessus. On décidera du centile à utiliser selon que l'on veut une estimation de l'erreur-type plus ou moins prudente. Par exemple, "100" nous donnera une estimation très prudente et "75", une estimation qu'il l'est moins.

4. BIBLIOGRAPHIE

1. ARORA, H.R. and BRACKSTONE, G.J., 'An Investigation of the Properties of Raking Ratio Estimators: ll with Cluster Sampling', Survey Methodology, Vol. 3, No. 2, 1977.

2. BANKIER, M., 'Power Allocations: Determining Sample Sizes for Subnational Areas', American Statistician, August 1988.

3. BANKIER, M., 'An Efficient Method of Estimating Raking Ratio Variance Under Stratified Cluster Sampling', rapport interne, DMES, Octobre 1987.

4. BRACKSTONE, G.J. and RAO J.N.K., 'An Investigation of Raking Ratio Estimators', Sankhya, Ser. C, Part 2, 41, 97-114, 1979.

5. COCHRAN, W.G., 'Sampling Techniques', Third Edition, John Wiley and Sons, 1977.

6. WOLTER, K.M., 'Introduction to Variance Estimation', Springer-Verlag, 1985.

La possibilité d'estimer le taux d'erreur nette de couverture lors du recensement de la population de 1991

Claude Julien[1]

1. INTRODUCTION

Le but premier d'un recensement a toujours été et demeure encore le dénombrement de la population vivant dans chaque région géographique d'un pays. Il doit inventorier cette population et ses conditions de vie. C'est la seule source de renseignements qui permet aux Canadiens d'obtenir une image complète et fiable de leur communauté et de leur pays.

Dans un projet de grande envergure tel le recensement du Canada, les résultats ne sont jamais parfaits. Même si l'on déploie des efforts considérables pour maintenir des normes élevées de qualité, des erreurs sont introduites au cours des étapes de la collecte et du dépouillement des données. Bien qu'il ne soit pas nécessaire que les statistiques soient parfaites pour être utiles, il est important que les utilisateurs aient conscience de la nature et de l'ampleur des erreurs que peuvent renfermer les données du recensement ainsi que des risques de baser des conclusions ou des décisions sur ces données. Afin d'informer les utilisateurs de données, il existe à Statistique Canada des programmes d'évaluation de la qualité des données du recensement. Un de ces programmes mesure l'erreur de couverture.

L'objectif principal de ce programme est d'estimer le taux d'erreur nette de couverture, c'est-à-dire le taux de sous-dénombrement ajusté en tenant compte du surdénombrement. Le second objectif de ce programme est d'estimer les diverses composantes de l'erreur nette, nommément le sous-dénombrement et le surdénombrement de ménages complets ou d'individus à l'intérieur des ménages.

---

1 Claude Julien est méthodologiste à Statistique Canada, Section de la qualité des données du recensement, Ottawa, Ontario.

Depuis 1966, Statistique Canada emploie la méthodologie de la Contre-vérification des dossiers (CVD) pour produire les estimations officielles du nombre et des caractéristiques des personnes et ménages manqués lors du recensement de la population. Pour la première fois en 1986, le recensement a donné lieu à une étude dont les objectifs étaient d'acquérir les connaissances et l'expérience nécessaires pour mesurer le surdénombrement tout en vérifiant que le taux de surdénombrement était petit comparativement au taux de sous-dénombrement. Cette étude a fourni des résultats intéressants (STATISTIQUE CANADA, 1990) qui ont mené Statistique Canada à considérer la possibilité d'effectuer dans l'avenir une seule enquête qui permettrait de réaliser tous les objectifs du programme d'évaluation de la couverture. Par contre, il est important de noter que la CVD fournira la mesure principale et officielle du taux de sous-dénombrement du recensement de 1991, alors que l'enquête considérée permettra alors d'évaluer et de compléter les résultats de la CVD.

Pour évaluer la possibilité d'effectuer une telle enquête, nous avons considéré deux options qui diffèrent principalement selon la base de sondage employée. L'une emploie une base de liste, tandis que l'autre fait appel à une base d'aires géographiques. Des problèmes inhérents à chacune des options, ajoutés au fait que la CVD sera encore, et avec raison, la source officielle pour estimer le sous-dénombrement, nous ont amené à conclure qu'une telle enquête serait trop coûteuse et risquée en termes de qualité des résultats. Pour 1991, nous avons opté pour une enquête plus simple, plus sûre et plus économique dont l'objectif sera de mesurer le taux de surdénombrement avec plus d'exactitude et de précision qu'en 1986.

Dans cet article, nous présenterons les résultats de cette évaluation. Pour commencer, nous discuterons de l'utilité d'effectuer une seule enquête pour estimer le taux d'erreur nette de couverture. Pour ce faire, nous définirons le surdénombrement pour ensuite en énumérer les diverses causes et les différentes méthodologies employées pour l'estimer. Ensuite nous décrirons les deux méthodologies qu'on a considérées pour l'enquête auprès des ménages et nous poursuivrons avec l'évaluation de ces deux méthodologies. Pour ce faire, nous discuterons des critères les plus importants, nous présenterons les résultats de l'évaluation et nous expliquerons pourquoi aucune des options n'a été retenue. Enfin, nous terminerons par la description de la méthodologie choisie.

2. DÉFINITIONS, CAUSES ET MÉTHODOLOGIES

Avant d'aborder l'évaluation de la possibilité d'une enquête qui permettrait d'estimer le taux d'erreur nette de couverture, il est important de mentionner pourquoi on voudrait employer une telle enquête. D'une part, la CVD, comme toute autre mesure de couverture, est sujette à certaines limites (BURGESS, 1988) surtout en ce qui concerne le sous-dénombrement des ménages. De plus, elle fournit peu d'information sur les causes du sous-dénombrement. D'autre part, on ne peut déceler et estimer certains types de surdénombrement qu'avec une enquête post-censitaire auprès des ménages. Dans cette section, nous définissons ce qu'est le surdénombrement et nous décrivons les différentes méthodologies employées pour l'estimer.

Dans le contexte du recensement de la population, un surdénombrement se produit lorsqu'une personne de la population cible est énumérée plus d'une fois, ou qu'une personne exclue de la population cible est énumérée au moins une fois. Les causes du surdénombrement sont nombreuses. Il est important de les énoncer car chaque type de surdénombrement requiert, pour sa mesure, une méthodologie différente.

Il y a d'abord des personnes dénombrées à la fois dans un logement privé et dans un logement collectif tel un hôpital ou une prison. Une enquête auprès des logements collectifs permet de mesurer ce type de surdénombrement. Les dossiers administratifs disponibles dans ces institutions fournissent des adresses supplémentaires où les résidents permanents pourraient aussi être dénombrés. La vérification des questionnaires du recensement correspondant à ces adresses permet de déterminer s'il y a eu surdénombrement.

Ensuite, il existe des personnes ou des ménages complets dénombrés plus d'une fois lorsque deux questionnaires du recensement sont distribués au même logement, soit par le même recenseur (mauvaise classification du logement), soit par deux recenseurs différents (mauvaise interprétation des limites du secteur de dénombrement). Ce type de surdénombrement apparaît dans une région géographique assez petite soit à l'intérieur d'un SD ou entre deux SD contigus. Puisqu'un SD contient au maximum 300 ménages il est possible de dépister ce type de surdénombrement en traitant l'information disponible sur la base de données du recensement à l'aide de programmes informatiques.

Il y a aussi des personnes ou des ménages complets dénombrés dans deux logements privés soit à cause d'un déménagement, soit lorsqu'un des logements est une résidence secondaire ou qu'un des membres du ménage travaille ou étudie ailleurs. Enfin, des personnes non éligibles fictives tels des visiteurs de l'étranger ou des enfants nés après la date du recensement ou des personnes fictives sont dénombrées lors du recensement. Ces deux derniers types de surdénombrement ne peuvent être dépister qu'en effectuant une enquête auprès des ménages. Pour effectuer cette enquête, Statistique Canada a considéré deux options.

3. LES OPTIONS

La première option considérée se veut une amélioration de l'étude du surdénombrement de 1986. Cette option permettrait d'estimer le surdénombrement de ménages complets et d'invididus ainsi que le sous-dénombrement d'individus.

La seconde option est une enquête semblable au POST ENUMERATION SURVEY effectué aux États-Unis (CHILDERS, 1987). Cette option permettrait d'estimer le taux d'erreur nette de couverture.

Les deux options consiste à employer un plan de sondage à deux degrés. Pour les deux options, les unités primaires d'échantillonnage sont des secteurs de dénombrement (SD), c'est-à-dire le secteur dénombré par un recenseur défini selon le nombre de ménages et les limites géographiques. Les deux options diffèrent par la suite selon la base de sondage employée pour sélectionner les unités secondaires.

Pour la première option on sélectionne des adresses inscrites aux registres de visites (RV) par le recenseur. L'ensemble des RV constitue une base de liste. Pour la seconde, on sélectionne des aires géographiques telles des îlots ou des segments clairement définis à l'intérieur du SD. L'ensemble des SD constitue une base géographique ou aréolaire. Des exemples de registre de visites (annexe A) et de carte de la géographie du recensement (annexe B) sont fournis à la fin du document.

La collecte des données consiste à énumérer des personnes et, pour la deuxième option, également des logements. Cette opération s'effectue indépendamment de la collecte déjà faite lors du recensement.

Des interviewers autres que ceux du recensement sont assignés à visiter chaque logement dans l'échantillon. À chacun de ces logements, l'interviewer dresse une liste de tous les occupants. En même temps, il pose des questions visant à confirmer l'éligibilité au recensement de chaque occupant et à obtenir des adresses supplémentaires où ceux-ci auraient pu être dénombrés lors du recensement. Les personnes ainsi énumérées forment l'échantillon-P, c'est-à-dire un échantillon de personnes.

Ensuite, l'information rapportée lors du recensement à chacun de ces logements est obtenue. Les personnes qui y sont listées forment l'échantillon-E, c'est-à-dire un échantillon d'énumérations.

L'échantillon-P contient des personnes qui auraient dû être dénombrées lors du recensement. Pour chaque personne on doit déterminer si elle a été dénombrée ou manquée. Les personnes manquées permettent d'estimer le sous-dénombrement.

L'échantillon-E contient des énumérations obtenues lors du recensement. Pour chaque énumération on doit déterminer si elle est correcte ou erronée. Les énumérations erronées servent à estimer le surdénombrement.

La détermination du statut de chaque personne et énumération s'effectue à l'aide de deux opérations. D'une part, les deux échantillons sont appariés. Puisqu'ils représentent les mêmes logements énumérés à deux moments différents, le taux d'appariement est élevé. Ceci permet de concentrer les ressources à résoudre les cas de non-appariement qui représentent un potentiel réel d'erreur de couverture. D'autre part, les questionnaires du recensement rapportés aux adresses supplémentaires fournies lors des entrevues sont vérifiés. Les personnes énumérées lors du recensement à la fois au logement choisi et à une adresse supplémentaire fournie sont dites surdénombrées.

Ensuite, les cas de non-appariement sont suivis, soit par réentrevue auprès des personnes concernées, soit par dépistage à l'aide de dossiers administratifs. Ce suivi a pour but de réconcilier les non-appariements ou de déterminer si des personnes ont été manquées ou énumérées par erreur. Les personnes manquées et les énumérations erronées ainsi trouvées servent à estimer l'erreur nette de couverture.

4. ÉVALUATION DES OPTIONS

Les critères qui ont servi à évaluer les deux options sont, entre autres, l'accessibilité et les coûts des bases de sondage considérées (registre de visites et cartes géographiques), l'utilité et la qualité de l'information qu'elles possèdent, ainsi que l'impact de leur utilisation sur les opérations et les résultats de l'enquête.

La base de sondage employée affecte principalement la période de collecte et l'appariement des échantillons P et E. Une période de collecte hâtive, c'est-à-dire près de la date du recensement donc en pleine période des vacances, diminue le taux de réponse car plus de ménages sont absents. Une période tardive augmente le nombre de ménages qui ont changé de logements depuis le recensement et, par conséquent, augmente le nombre de non-appariements. La base de sondage désirée doit permettre un appariement efficace, pour diminuer le nombre de suivis, et rapide, pour accorder le temps nécessaire pour effectuer un suivi adéquat et complet.

Selon ces critères, les principaux arguments en faveur de l'utilisation des registres de visites sont nombreux. Premièrement, les RV fournissent une base de sondage accessible et peu coûteuse. En fait, ils proviennent des opérations du recensement et ne nécessitent aucun coût de développement additionnel. Cette base fournit de l'information utile qui indique la possibilité d'erreur de couverture. En 1986, on a observé que le type de questionnaire rempli, long ou court, et l'indicateur du retour par la poste (colonnes 2 et 11 du RV à l'annexe A) avaient une incidence sur le surdénombrement. Cependant, en raison du faible taux de surdénombrement de 0.2%, soit 58 personnes parmi les 31 371 personnes choisis, ces

observations n'ont qu'une nature qualitative et n'ont aucune justification empirique. Enfin, la base des RV fournit un lien direct avec le questionnaire du recensement qui facilite grandement l'appariement.

Par contre, les registres de visites présentent aussi beaucoup d'inconvénients. La qualité des adresses et de l'information inscrites est quelquefois douteuse, surtout en ce qui concerne l'indicateur du retour par la poste. De plus, dans les SD ruraux l'adresse designe souvent plus d'un logement, par exemple route 307 ou RR 2. Étant donné que la collecte de l'enquête et celle du recensement sont des opérations indépendantes, ces ambiguïtés occasionnent des erreurs de collecte lorsque le ménage contacté lors de l'enquête diffère de celui sélectionné. L'impact d'une telle erreur est amplifiée par le fait qu'elle n'apparaît que lors de l'appariement, beaucoup trop tard pour être corrigée adéquatement. Pour éviter partiellement de tels problèmes, on doit attendre que les RV soient de bonne qualité et le plus complet possible. Malheureusement, ceci repousse la sélection et la collecte à la fin du mois de juillet et risque d'augmenter le taux de non-appariements. Enfin, cette méthodologie ne permet de dépister que le sous-dénombrement d'individus et non celui des logements, et, par conséquent, ne satisfait pas tous les objectifs du programme d'évaluation.

Les avantages militant en faveur de la base géographique sont aussi nombreux que ceux en faveur de la base de registres. Premièrement, cette base est indépendante des opérations du recensement. Ceci permet d'effectuer la sélection de l'échantillon et l'énumération des logements un mois ou deux avant le recensement. Ainsi, on peut assurer la qualité de l'échantillon et obtenir une description physique des logements à visiter après le recensement. Cette qualité est particulièrement attrayante dans les SD ruraux. Suite à l'énumération des logements, une liste de logements à contacter est produite et la collecte des données peut commencer dès le début du mois de juillet, c'est-à-dire immédiatement après la fin de la collecte du recensement. Cette base a aussi l'avantage de fournir un échantillon de logements contigus qui rend l'appariement plus robuste, moins vulnérable aux erreurs dues à l'ambiguïté de certaines adresses. Enfin, cette base permet de dépister le sous-dénombrement de logements et, par conséquent, satisfait tous les objectifs du programme d'évaluation.

Mais, tout comme la base des RV, cette base présente de sérieux inconvénients. Tout d'abord, elle n'est disponible que pour les SD des villes de 100 000 habitants et plus. Pour ces SD on

peut produire automatiquement une carte où chaque îlot est indiqué (voir un exemple à l'annexe B). Pour les autres SD, il faut manuellement délimiter des îlots ou des segments à partir de plans de ville ou de cartes topographiques. Cette opération requiert de coûts de développement considérables. Ensuite, elle ne fournit pas de lien direct avec l'information rapportée lors du recensement. Cet inconvénient, peut-être le plus important, complique grandement l'identification de l'échantillon-E. Bien que ceci soit possible dans les SD urbains, car le recenseur doit inscrire le numéro d'îlot au registre de visites (colonne 1 à l'annexe A), ceci est presque impossible pour les SD ruraux où le concept d'îlot n'existe pas. Ainsi, beaucoup d'opérations manuelles sont requises pour développer cette base, sélectionner l'échantillon d'îlots et identifier l'échantillon-E. Chacune de ces opérations est sujette à des erreurs qui peuvent affecter la qualité des résultats. Enfin, cette méthodologie emploie essentiellement la même méthode de collecte qu'au recensement. Ceci entraîne une corrélation entre la probabilité de manquer un logement lors du recensement et la probabilité de le manquer également lors de l'enquête. Malheureusement, cette situation occasionne une sous-estimation du sous-dénombrement des logements.

Après avoir pesé les avantages et les inconvénients de chaque option, nous avons conclu que ni l'une ni l'autre était acceptable. La première option implique des problèmes lors de la collecte et une période de suivi trop tardive. En plus, elle ne satisfait pas tous les objectifs du programme. La seconde est coûteuse, ne permet pas de clairement identifier l'échantillon-E et risque de fournir des résultats de qualité assez faible.

5. L'OPTION CHOISIE

Dans le contexte actuel, le projet d'estimer le taux d'erreur nette avec une seule enquête a été abandonné et, finalement, une troisième option, moins ambitieuse mais plus sûre, a été acceptée.

Tout comme la première option, celle-ci consiste en la sélection d'un échantillon d'adresses à partir des registres de visites. Au lieu de transcrire seulement l'adresse et le nom d'un des occupants (colonnes 2, 3 et 4 du RV à l'annexe A) comme ce fut le cas en 1986, on transcrit sur le questionnaire de l'enquête l'adresse et le numéro de téléphone du ménage ainsi que

le nom et certaines caractéristiques (sexe, date de naissance, etc.) de chaque personne listée sur le questionnaire du recensement. Du bureau régional, l'interviewer contacte par téléphone les ménages choisis. L'interview s'effectue en personne, à l'adresse transcrite, seulement lorsque le numéro de téléphone n'est pas rapporté lors du recensement. Tout comme les deux options considérées précédemment, l'interviewer pose des questions pour confirmer l'éligibilité au recensement de chaque personne transcrite et obtenir des adresses supplémentaires où elles auraient pu être énumérées lors du recensement.

Au lieu d'utiliser les ressources pour l'appariement et la réconciliation des non-appariements, celles-ci sont allouées pour augmenter le taux de réponse et pour mieux dépister les personnes transcrites qui ont déménagé. Pour ces personnes, il faut employer tous les moyens disponibles pour trouver leur nouvelle adresse et les contacter. En cas d'échec, il faut au moins confirmer leur existence et leur éligibilité en contactant soit les occupants actuels du logement choisi, les voisins ou tout autre source d'information appropriée.

Plusieurs avantages accompagnent cette méthode. Premièrement, elle réduit considérablement les coûts de collecte. On estime que plus de 80 % des interviews s'effectueront par téléphone. Elle ne nécessite aucun appariement. Elle emploie des opérations déjà éprouvées par d'autres enquêtes. Par exemple, la sélection de l'échantillon et la transcription se feront en empruntant l'expérience acquise lors des enquêtes post-censitaires; l'interview et le dépistage s'effectueront de façon assez semblable à ce qui se fait déjà lors des opérations de la Contre-vérification des dossiers.

Par contre, cette méthodologie ne permet d'estimer que le surdénombrement. Elle nécessite des ressources pour contrôler la qualité de la sélection et la transcription des noms et adresses. Enfin, bien qu'une entrevue de ce genre ait déjà été effectuée pour la Contre-vérification des dossiers, une telle entrevue dérange les gens qui pense qu'on évalue la qualité de l'information qu'ils ont rapportée lors du recensement. Il est donc essentiel de bien présenter le but de l'enquête au début de l'entrevue.

6. CONCLUSION

Dans le contexte actuel où le sous-dénombrement a été estimé par la Contre-vérification des dossiers depuis 1966, et qu'il en sera de même en 1991, l'idée d'effectuer une seule enquête pour estimer le taux d'erreur nette de couverture a été abandonnée.

Bien qu'une telle enquête existe aux États-Unis et ailleurs, cette décision se justifie par le fait qu'une telle enquête coûterait cher, surtout en termes de ressources et peut-être même en termes de qualité. De plus, le besoin premier pour 1991 est de produire une mesure fiable du surdénombrement pour complémenter les estimations de sous-dénombrement de la Contre-vérification des dossiers.

Pour satisfaire ce besoin, on effectuera une enquête dont le seul but sera d'estimer le nombre de personnes énumérées par erreur. Cette enquête sera beaucoup plus simple, et par le fait même plus sûre en termes de qualité des résultats, et plus économique que les deux options étudiées.

BIBLIOGRAPHIE

BURGESS, R. (1988). Évaluation des estimations du sous-dénombrement obtenus par la contre-vérification des dossier du recensement du Canada. Techniques d'enquêtes. Volume 14, numéro 2, 147-168.

CHILDERS, D., DIFFENDAL,G., HOGAN, H., SCHENKER, N. and WALTER, K. (1987). The technical feasibility of correcting the 1990 Census. Proceedings of the Social Statistics Section, American Statistical Association. 36-45

STATISTIQUE CANADA, Guide à l'Intention des Utilisateurs sur la Qualité des Données du Recensement de 1986: Couverture. Publication 99-145F

ANNEXE A

SECTION I — LOGEMENTS PRIVÉS OCCUPÉS — SECTION I

MAISONS INDIVIDUELLES, APPARTEMENTS, ETC.

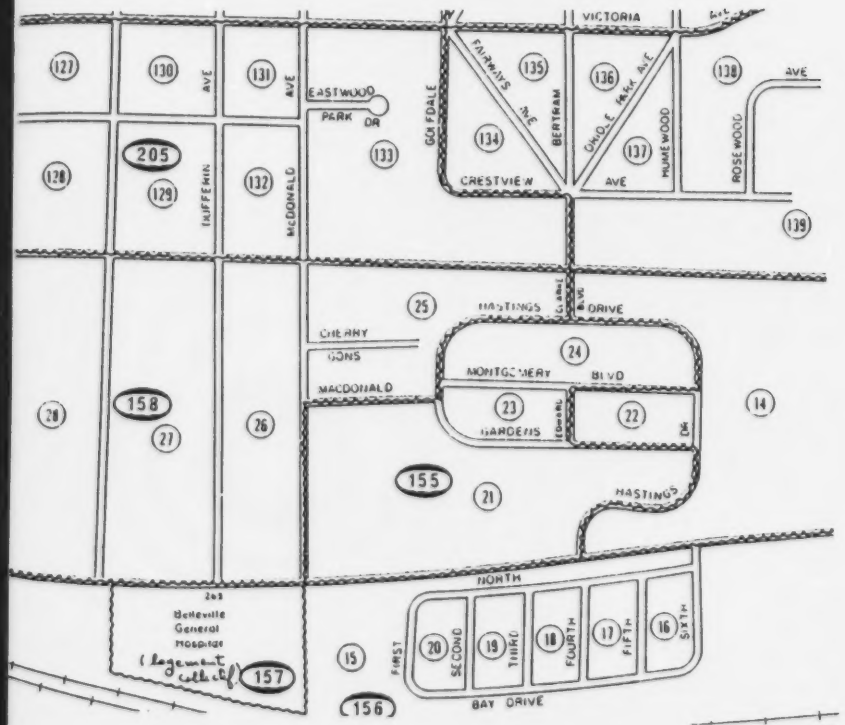
LOCALITÉ NON CONSTITUÉE ET/OU ÎLOT N° (si nécessaire)	MÉNAGE N°	SITUATION EXACTE DU LOGEMENT 1) Pour les cités, villes et villages, inscrivez le nom de la rue et le numéro de voirie ainsi que le numéro d'appartement (s'il y a lieu). 2) Dans les autres cas, indiquez : a) le lot, le concession et le numéro de la route rurale (Québec et Ontario); b) la section, le township ou canton, le rang et le méridien (Prairies); c) la paroisse, le township ou canton, etc. (autres provinces). 3	NOM DE LA PERSONNE 1 DE CE MÉNAGE Nom, prénoms et initiales (en lettres moulées) La Personne 1 est une des personnes suivantes : – l'un des conjoints (épouse, épouse) de tout couple marié demeurant ici; – l'un des deux partenaires en union libre; – le père ou la mère lorsque seulement un des deux vit avec un ou plusieurs de ses enfants célibataires (jamais mariés) quel que soit leur âge. Si aucune de ces catégories ne s'applique, choisissez n'importe quel membre adulte du ménage comme Personne 1. 4	DATE DE LA PREMIÈRE VISITE À CE MÉNAGE 5 Date	NOMBRE DE RÉSIDENTS HABITUELS 6	LANGUE DE PRÉFÉRENCE DU MÉNAGE 7 Ang. Franç.
		N° de voirie / Rue / App. n°				
	031					
	032					
	033					
	034					
	035					
	036					
	037					
	038					
	039					
	040					
	041		→ Spécial recensement court			
	042					
	043					
	044					
	045		← spécial recensement long			
	046					
	047					
	048					
	049					
	050					
	051					
	052					

LOGEMENTS DE PLAIN-PIED, ETC.

MÉNAGE N°	UN DES RÉSIDENTS EXPLOITE-T-IL UNE EXPLOITATION AGRICOLE OU UNE PARTIE DE CELLE-CI OU PLUS AU TITRE DE LA VENTE DE PRODUITS AGRICOLES AU COURS DES 12 DERNIERS MOIS?	QUESTIONNAIRE AGRICOLE N°	RETOUR PAR LA POSTE SEULEMENT DATE QUESTIONNAIRE REÇU PAR LA POSTE	COMPLÉTÉ PAR SUIVI TÉLÉPHONIQUE	COMPLÉTÉ PAR SUIVI SUR PLACE	NOMBRE DE RÉSIDENTS TEMPORAIRES	MÉNAGES – TRAVAIL TERMINÉ				EXPLOITATIONS AGRICOLES – TRAVAIL TERMINÉ			REMARQUES
	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
		Oui / Non		2A / 2B	Date	Date	3	2A	2B	4A	2A/4A	2B/4B	6	4A	6/4B	
031																
032																
033																
034																
035																
036																
037																
038																
039																
040																
041																
042																
043																
044																
045																
046																
047																
048																
049																
050																
051																
052																

ANNEXE B

G86-14 CARTE DE SECTEURS DE DÉNOMBREMENT URBAINS



Enumeration Area Boundary with number ..... Limite de secteur de dénombrement avec le numéro ... 155
Federal Electoral District Boundary with number ..... Circonscription électorale fédérale avec le numéro 35089
City Block Number ..... Numéro d'îlot 21

Urban EA - G 86-14
SD Urbain - G 86-14

G86-12 CARTE DES SECTEURS DE DÉNOMBREMENT RURAUX

CENSUS RECENSEMENT CANADA

Census Division
Census Subdivision
Enumeration Area Boundary
Enumeration Area Number
Incorporated City Town or Village
Federal Electoral District

Division de recensement
Subdivision de recensement
Limite de secteur de dénombrement
Numéro de secteur de dénombrement
Cité ville ou village constitué
Circonscription électorale fédérale

Geography Division
Division de la Géographie

Rural EA - G 86-12
SD Rural - G 86-12

STRATOR, un logiciel généraliste pour le traitement de données d'enquête

S. LASSARRE (INRETS-DERA, ARCUEIL, FRANCE),
J. P. ROWENCZYK (DEMOSCOPIE, PARIS, FRANCE)

Résumé

A l'origine, créé pour des besoins d'estimation de trafic sur les routes, STRATOR est actuellement un logiciel généraliste permettant de traiter les données issues d'enquêtes aléatoires utilisant des plans de sondage complexes, pour lesquels l'utilisateur peut mettre en oeuvre jusqu'à 5 degrés de tirage et 5 niveaux de stratification. STRATOR calcule les estimateurs souhaités (somme dilatée, quotient)ainsi que leur précision. La présentation de STRATOR et de ses applications sera faite à partir de cas concrets.

I - Historique

STRATOR a été développé en 1988 par le Département d'Evaluation et de Recherche en Accidentologie de l'Institut National de Recherche dans les Transports et leur Sécurité et DEMOSCOPIE pour les besoins de l'Observatoire National de la Sécurité Routière, qui est un service de la Direction de la Sécurité et de la Circulation Routière (Lassarre, 1989).

Pour apprécier la variance de l'estimateur du parcours annuel France entière sur le réseau des chemins départementaux, l'Observatoire a lancé en juin 1987 un appel d'offres pour la réalisation d'un logiciel capable de calculer la précision d'un plan de sondage complexe, suite auquel DEMOSCOPIE a été retenu.

Bien que ce logiciel fût conçu pour une application précise, les développeurs ont cherché à le rendre le plus universel possible par une capacité de traitement d'une grande variété de plans de sondage.

II - Spécificités générales de STRATOR

STRATOR calcule la variance d'estimateurs d'une fonction d'une variable d'intérêt Y, qui prend ces valeurs sur une population d'individus, à partir des mesures disponibles sur un échantillon issu d'un plan de sondage complexe.

Les fonctions de la variable peuvent être :

- soit un total T (Y),
- soit un ratio $R(Y_1, Y_2) = T(Y_1)/T(Y_2)$,

La variable Y peut être une variable quantitative, ou une indicatrice d'un caractère qualitatif, notée 1_C qui vaut 1 si l'indicateur possède cette caractéristique, 0 sinon.

La ratio contient comme cas particulier la proportion avec $Y_1 = 1_C$ et $Y_2 = 1_P$, variable indicatrice de l'appartenance à la population.

La complexité d'un plan de sondage gérée par STRATOR s'étend jusqu'à une combinaison de dix degrés de sondage qui peuvent être stratifiés ou sous-stratifiés.

Dans chaque strate et unité, STRATOR ne fonctionne que sur le mode du sondage aléatoire simple sans remise SASSR (Deville, 1987), dit plan à probabilité égale avec remise PESR (Grosbras,1987).

Si la population est de grande taille, ou si la taille de l'échantillon est petite devant la taille de la population, un SASSR est assimilable à un sondage aléatoire simple avec remise dit SASAR ou PEAR .

Dans STRATOR, les formules de variance font appel au facteur correctif 1-n/N avec N et n les tailles de la population et de l'échantillon. Il devient négligeable lorsque le taux de sondage n/N est faible. Une application de cette formule est de rendre nulle la variance d'une strate d'unités inclus d'office dans l'échantillon sans tirage au hasard suite à l'égalité des tailles de population et d'échantillon.

Deux estimateurs sont disponibles pour le total :

- l'estimateur somme dilatée
- l'estimateur par le quotient.

L'estimateur de base est du type somme dilatée:

$$\hat{T}(Y) = \frac{N}{n} \sum_{i=1}^{n} y_i$$

avec n et N la taille de l'échantillon et de la population, et y_i la mesure de la variable d'intérêt sur l'échantillon.

Le second à être privilégié est l'estimateur du quotient calculé à l'aide du total T(X) de la mesure d'une variable auxiliaire X selon la formule:

$$\hat{T}^q(Y) = T(X) \frac{\frac{N}{n} \sum_{i=1}^{n} y_i}{\frac{N}{n} \sum_{i=1}^{n} x_i} = T(X) \frac{\hat{T}(Y)}{\hat{T}(X)}$$

L'estimateur du ratio est naturellement du type quotient

$$R(Y_1, Y_2) = \frac{\hat{T}(Y_1)}{\hat{T}(Y_2)}$$

II.1. Principes de structuration de la base de sondage

Une base de sondage est interprétée et décrite comme une hiérarchie de parties de l'ensemble de la population (Benzécri,1973). La racine ou le sommet de la hiérarchie est l'univers ou la population dans son ensemble notée U_0, qui englobe tous les individus. L'individu, formant l'unité de sondage de plus bas niveau, appartient à l'ensemble des éléments terminaux de la hiérarchie.

"Un plan de sondage est dit complexe s'il n'est pas de type aléatoire simple et s'il utilise une information auxiliaire individuelle connue sur l'ensemble de la base de sondage" (Deville,1987). L'information auxiliaire traitée par STRATOR se présente sous deux formes :

- l'unité qui peut être une strate,
- la variable auxiliaire X.

Pour l'individu de l'échantillon, sont connus avant l'enquête proprement dite, son appartenance à des unités ou des strates, et la valeur d'une variable auxiliaire.

La complexité d'un plan de sondage gérée par STRATOR s'étend jusqu'à une combinaison de dix degrés de sondage qui peuvent être stratifiés ou sous-stratifiés, représentée par cinq couples strate/unité notés S/U :

$$U_0\ S_1\ U_1\ S_2\ U_2\ S_3\ U_3\ S_4\ U_4\ S_5\ U_5$$

avec U_0 pour la population totale ou l'univers. Le niveau de stratification est considérée comme un degré, puisque "un tirage stratifié est un plan à deux degrés où le tirage en premier degré est un recensement" (Gourieroux,1981).On emploie le terme sous-stratification lorsque la division en strate ne peut pas être opérée sur la population entière mais sur des parties de celle-ci appartenant à des unités de degré inférieure au sommet U_0 (Desabie,1971).

II.2. Principes de calcul de l'estimateur et de sa variance

L'analyse des formules de l'estimateur de la somme dilatée fait apparaître un opérateur noté S qui, appliqué aux estimateurs des totaux des nœuds $\{a_{k+1}\}$ du niveau k+1, fournit les estimateurs des totaux des nœuds $\{a_k\}$ au niveau supérieur k tel que:

$$\hat{T}_{a_k}(Y) = S\ \{\ \hat{T}_{a_{k+1}}(Y)/a_{k+1}\ E\ Sui(a_k)\ \}$$

où $Sui(a_k)$ = l'ensemble des successeurs immédiats de a_k.

Si le niveau k est le degré d'une unité,

$$\mathbb{S} = \frac{N}{n} \int_{i=1}^{n} \qquad \text{avec } n=\text{Card}\{\text{Sui}(a_k)\},$$

N=taille de la population au nœud a_k.

Si le niveau k est le degré d'une stratification,

$$\mathbb{S} = \sum_{h=1}^{H} \qquad \text{avec H le nombre de strates au nœud } a_k.$$

Au niveau le plus bas, pour a_k appartenant à Term A, l'ensemble des éléments terminaux de l'arbre, $\hat{T}_{a_k}(Y)$ est égal à la mesure de la variable d'intérêt sur l'unité de plus bas degré.

Pour un sondage à deux degrés avec stratification au premier degré, l'estimateur du total s'écrit:

$$\hat{T}_{a_0}(Y) = \sum_{h=1}^{H} \frac{M_h}{m_h} \int_{i=1}^{m_h} \frac{N_{hi}}{n_{hi}} \int_{j=1}^{n_{hi}} y_{hij}$$

La formule de la variance prend aussi une forme récursive:

$$\text{var}(\hat{T}_{a_k}(Y)) = \mathbb{S}_1\{\frac{N}{n-1}\frac{N-n}{N}\text{résidu}^2(\hat{T}_{a_{k+1}}(Y))/a_{k+1} \text{ E Sui}(a_k)\}$$

$$+\mathbb{S}_2\{\text{var}(\hat{T}_{a_{k+1}}(Y))/a_{k+1} \text{ E Sui}(a_k)\}$$

Si le niveau k est le degré d'une unité:

$$\mathbb{S}_1 = \mathbb{S}_2 = \frac{N}{n} \int_{i=1}^{n}$$

Si le niveau k est le degré d'une stratification:

$$\mathbb{S}_1=0 \text{ et } \mathbb{S}_2 = \sum_{h=1}^{H}$$

$$\text{résidu}(\hat{T}_{a_{k+1}}(Y)) = \hat{T}_{a_{k+1}}(Y) - \frac{\hat{T}_{a_k}(Y)}{N}$$

La variance du niveau le plus bas est égale à zéro.

Pour l'exemple précédent, la variance s'écrit:

$$\text{var}(\hat{T}_0(Y)) = \sum_{h=1}^{H} \frac{M_h}{m_h}\left[\sum_{i=1}^{m_h} M_h \frac{M_h - m_h}{M_h(m_h-1)}\left(y_{hi} - \frac{y_h}{M_h}\right)^2 + \frac{N_{hi}}{n_{hi}}\left\{\sum_{j=1}^{n_{hi}} N_{hi}\frac{N_{hi}-n_{hi}}{N_{hi}(n_{hi}-1)}\left(y_{hij} - \frac{y_{hi}}{N_{hi}}\right)^2\right\}\right]$$

Structurellement, l'estimateur par le quotient à un niveau k s'exprime en fonction des estimateurs directs des totaux de la mesure de la variable d'intérêt Y et de la mesure de la variable auxiliaire X :

$$\hat{T}_{a_k}^{q}(Y) = T_{a_k}(X)\frac{\hat{T}_{a_k}(Y)}{\hat{T}_{a_k}(X)}$$

Le niveau où peut être calculé l'estimation par le quotient dépend de la disponibilité du total de la variable auxiliaire à ce niveau, noté T(X).

La variance de l'estimateur du quotient se calcule par une formule récursive du même type que pour l'estimateur direct en remplaçant $\hat{T}_{a_{k+1}}(Y)$ par la variable transformée $\hat{Z}_{a_{k+1}}(Y)$ (Deville et Roth,1986):

$$\hat{Z}_{a_{k+1}}(Y) = \hat{T}_{a_{k+1}}(Y) - \hat{R}\,\hat{T}_{a_{k+1}}(X)$$

avec $\hat{R}$ l'estimateur du ratio au niveau où opère le quotient:

$$\hat{R} = \frac{\hat{T}_{a_k}(Y)}{\hat{T}_{a_k}(X)}$$

Pour assurer la validité de l'estimateur par le quotient, il faut apprécier l'importance du biais à l'aide d'une règle fondée sur la valeur du coefficient de variation de la mesure de la variable auxiliaire.

Si l'on effectue un quotient séparé au nœud a_k d'une strate d'une stratification comportant H strates, la règle s'écrit:

$$\sqrt{H}\,\frac{\sqrt{\mathrm{var}(\hat{T}_{a_k}(X))}}{\hat{T}_{a_k}(X)} < 0.2 \text{ (Desabie,1971) ou } 0.3 \text{ (Cochran,1977)}$$

Si l'on effectue le quotient au nœud a_k d'une unité de niveau k, la règle devient:

$$\frac{\sqrt{\mathrm{var}(\hat{T}_{a_k}(X))}}{\hat{T}_{a_k}(X)} < 0.1$$

III. Principes de fonctionnement

Ce logiciel tourne sur tout compatible IBM PC, XT ou AT équipé de 640 K0 de mémoire centrale, d'un lecteur et d'un disque dur, fonctionnant sous MS-DOS, avec un écran couleur, ou noir et blanc. Les impressions s'effectuent sur toutes les imprimantes compatibles IBM GRAPHIC, EPSON, HP LASER JET ... Il est nécessaire d'utiliser le mode compressé pour les imprimantes 80 colonnes.

L'information de base lue par STRATOR est le fichier des données du sondage où à un individu correspond un enregistrement composé de deux parties :

- un ensemble des champs décrivant les caractéristiques du plan de sondage, et remplis par les codes des strates et des unités des divers degrés,

- un ensemble de champs qui contiennent les valeurs des variables d'intérêt Y, et éventuellement des variables auxiliaires X, ou d' autres variables d'intérêt Y_2.

Les tailles de population notées N des différents degrés doivent être connues, car c'est à partir d'elles que sont calculées les sommes dilatées ; ainsi que les totaux de la variable auxiliaire T(X) pour calculer l'estimateur du quotient.

Dans la terminologie de STRATOR, on manie trois sortes d'objets :

. le jeu de données qui est physiquement représenté par un fichier des individus d'un échantillon d'une population décrits par des caractéristiques de la population et une série de variables

. le plan ou la structure du sondage qui est représenté par la hiérarchie des degrés avec les couples strate/unité; y sont joints les tailles de population et les totaux de la variable auxiliaire ;

Exemples : $U_0\ S_1\ S_2\ S_3\ U_3$ pour un plan de sondage à un degré avec trois niveaux de stratification, $U_0\ S_1\ U_1\ U_2$ pour un plan de sondage à deux degrés avec stratification des unités primaires.

. la fonction à estimer et son estimateur qui peuvent être :

- un total estimé en somme dilatée,

- un total estimé par le quotient,

- un total estimé en somme dilatée + par le quotient,

- un ratio estimé par le quotient global,

à calculer sur un couple de variables présentes dans le fichier :

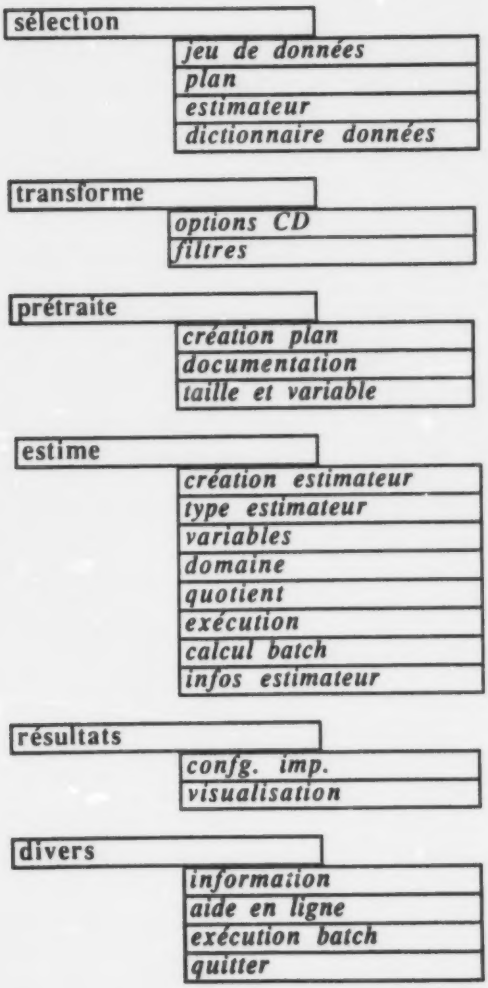
(Y ,.) pour un total,

(Y, X) pour un total par le quotient,

(Y_1, Y_2) pour un ratio.

III.1. Le menu et les fonctions

STRATOR, écrit en dBASE III, propose six menus principaux ; chacun déroulant une série de sous/menus :



A partir des six menus principaux de STRATOR, sont accessibles vingt trois fonctions.

Les menus principaux sont rangés dans l'ordre naturel des opérations à effectuer :

sélection : pour choisir un jeu de données, un plan de sondage et un estimateur dans différentes listes ;

transforme : pour transformer ou coder des variables quantitatives ou qualitatives (une transformation particulière pour le besoin du traitement des données du sondage sur les chemins départementaux est disponible) ;

prétraite : pour saisir la structure du plan de sondage, les tailles de population et les totaux de la variable auxiliaire ;

estime : pour lancer le calcul de l'estimateur après avoir choisi un estimateur et les variables d'intêret et auxiliaire dans une liste, défini un domaine et les niveaux où s'applique le quotient ;

résultats : pour visualiser et imprimer les résultats après le choix d'une configuration d'impression ;

divers : pour quitter, avoir accès à des informations, lancer une estimation en différé.

III.2 Application: estimation du parcours sur les chemins départementaux

Le jeu d'essai qui suit, est tiré de l'enquête 1984 de circulation sur les chemins départementaux (Lassarre, 1989). L'objectif est d'estimer le parcours annuel, exprimé en milliards de véhicules x kilomètres, à partir de comptages journaliers effectués sur un échantillon de sections de C.D. tirés au hasard suivant un plan de sondage complexe.

Nous comparerons deux estimateurs du parcours obtenus à partir de comptage du débit d_i sur une section de longueur l_i :

- l'estimateur somme dilatée par sondage aléatoire simple sans remise:

$$\hat{t}(LD) = \frac{N}{n} \sum_{i=1}^{n} l_i d_i$$

- l'estimateur du quotient par sondage aléatoire simple sans remise avec T(L) la longueur totale du réseau connue:

$$\hat{t}_q(LD) = T(L) \frac{\frac{N}{n} \sum_{i=1}^{n} l_i d_i}{\frac{N}{n} \sum_{i=1}^{n} l_i}$$

La structure hiérarchique du plan de sondage est celle d'un plan à deux degrés avec une sous-stratification du premier degré (Desabie,1971):

Uo	univers
U1	le département
S2	la catégorie de CD
U2	la section

avec une strate de dix départements comme univers, le département comme unité primaire, la catégorie de CD pour sous-stratifier, et la section comme unité secondaire.

Deux départements: la Charente Maritime (17) et la Haute Marne (52) ont été tirés dans la strate de dix départements. Le réseau de CD de chaque département est sous-stratifié en deux catégories d'après leur couleur sur la carte Michelin: rouge codée 1 et jaune codée 2. Quatre sections sont tirées dans chaque département par catégorie. Le parcours annuel est obtenu pour chaque section de l'échantillon à partir d'une série de comptages des véhicules circulant sur ces sections.

Le fichier des données du sondage appelé IN1752.SD, qui comporte 2x2x4, soit 16 individus, ici les section de C.D., a la structure suivante :

le numéro du département

la catégorie de CD

le numéro d'ordre de la section par catégorie et département

la longueur de la section

les mois et les jours du début et la fin de comptage

les six débits journaliers

le parcours annuel en milliers de véhicules x kilomètres

le nombre de Poids-Lourds et le nombre total de véhicules comptés sur une période de deux heures.

La variable d'intérêt est le parcours annuel pour la section de l'échantillon, et la variable auxiliaire la longueur de la section. Les tailles de population en terme de nombre de sections , et les totaux de la variable auxiliaire en terme de longueur sont donnés dans le tableau suivant.

Département	Catégorie	Taille	Mesure
		nombre de sections	longueur des CD km
17	1	25	573
17	2	59	936
17	total	84	1509
52	1	12	259
52	2	38	556
52	total	50	815

Tableau 1. Nombre de sections et longueur des CD par département et catégorie de CD.

IV. Conclusion

STRATOR est un logiciel interactif dédié à l'estimation de la variance des estimateurs somme dilatée et quotient d'un total ou d'une proportion pour un plan de sondage qui peut comporter jusqu'à dix degrés dont cinq de stratification.

Il travaille sur une variable à la fois. Une version multidimensonnelle,capable de traiter plusieurs variables conjointement comme le font les logiciels PC CARP (Schnell et al., 1988), et SESUDAAN (Marchand, 1988) est un développement futur envisagé.

La version actuelle du logiciel a tourné pour exploiter des données de sondage issues du champ des transports (Lassarre et al., 1990), mais par sa conception elle est apte à traiter des enquêtes de type probabiliste de tout domaine.

Bibliographie

J.P. Benzécri et Collaborateurs. L'Analyse des Données 1 La Taxinomie. Dunod. 1973. Paris.

W. Cochran. Sampling Techniques. Third edition. John Wiley & Sons. 1977. New York.

J. Desabie. Théorie et Pratique des Sondages. Dunod. 1971. Paris.

J.C. Deville, N. Roth. La précision des enquêtes sur l'emploi. Economie et Statistique n° 193-194, Novembre 1986, p 127-134.

J.C. Deville. Les sondages complexes in Les Sondages. J.J. Drœsbecke, B. Fichet, P. Tassi, éditeurs. Economica. 1987. Paris.

C. Gouriéroux. Théorie des sondages. Economica. 1981. Paris.

J.M. Grosbras. Méthodes statistiques des sondages. Economica. 1987. Paris.

S. Lassarre. Estimation par sondage du parcours sur les chemins départementaux. Rapport INRETS n° 103. 1989. Arcueil.

S. Lassarre. STRATOR version 1.3 : Présentation technique. Note INRETS. 1989. Arcueil.

S. Lassarre, J.P. Rowenczyk, C. Mohr, B. Caruette. STRATOR: un logiciel d'aide à l'exploitation des enquêtes par sondage sur le réseau routier.Route et Informatique. Presses de l'ENPC. 1990. Paris.

F. Marchand. Notice d'utilisation de SUSEDAAN. INED. 1988. Paris.

D. Schnell, W.J. Kennedy, G. Sullivan, H.J. Park, W.A. Fuller. Logiciel d'ordinateur personnel pour l'estimation de la variance dans les enquêtes complexes. Techniques d'enquête. Vol 14, n° 1, 1988, 63-73.

Correction simple pour le biais des estimateurs de variance traditionnels

par

Yves Leblond[1]

1 Introduction

Les estimateurs de variance traditionnels d'estimateurs non linéaires de caractéristiques d'une population sont généralement biaisés lorsque la taille d'échantillon est relativement petite. Par exemple, les estimateurs de variance obtenus par la méthode de linéarisation de Taylor et la méthode dite du jackknife ont respectivement tendance à sous-estimer et surestimer la variance des estimateurs habituellement utilisés en échantillonnage. Ceci est vrai en particulier pour l'estimateur par le quotient $\bar{X}\bar{y}/\bar{x}$.

Nous suggérons une méthode simple permettant de réduire le biais des estimateurs de variance habituellement utilisés pour l'estimateur par le quotient. Cette méthode consiste à obtenir un estimateur de variance modifié en multipliant l'estimateur de variance (jackknife ou Taylor) par un facteur de correction qui ne dépend que de la taille d'échantillon n et des valeurs de la variable auxiliaire x mais non des valeurs de y. Ce facteur de correction tend vers 1 lorsque n augmente et ainsi l'estimateur de variance modifié obtenu est asymptotiquement équivalent à l'estimateur traditionnel. Ce facteur est relativement simple à justifier et simple à calculer. Il est obtenu en utilisant une méthode assistée d'un modèle.

Nous exposerons dans la section 2 la méthode utilisée afin d'obtenir l'estimateur de variance modifié. Nous appliquerons cette méthode à trois estimateurs de variance traditionnels de l'estimateur par le quotient. Nous présenterons à la section 3 quelques résultats de simulations illustrant le bien fondé de notre méthode de réduction du biais.

2 Estimateur par le quotient

Considérons un plan d'échantillonnage aléatoire simple (SI) de n unités parmi N. Posons $f = n/N$ et définissons l'estimateur par le quotient de la moyenne de population $\bar{Y}$ par

$$\bar{y}_Q = \bar{X}\,\frac{\bar{y}}{\bar{x}}$$

où $\bar{X} = \sum_{i=1}^{N} \frac{x_i}{N}$, $\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$ et $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$.

---

1 Yves Leblond, Division des méthodes enquêtes-entreprise, Statistique Canada, Ottawa K1A 0T6.

On considère l'estimation de l'erreur quadratique moyenne (EQM), $EQM(\bar{y}_Q) = E(\bar{y}_Q - \bar{Y})^2$ où $E(\cdot)$ désigne l'espérance par rapport au plan SI. Considérons trois estimateurs de variance de $\bar{y}_Q$ couramment utilisés: l'estimateur obtenu par linéarisation de Taylor ($\hat{V}_T$), l'estimateur de Taylor avec correction de Royall ($\hat{V}_{T2}$) et l'estimateur jackknife ($\hat{V}_J$) définis respectivement par

$$\hat{V}_T = \frac{1-f}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{R}x_i)^2}{n-1}$$

$$\hat{V}_{T2} = \left(\frac{\bar{X}}{\bar{x}}\right)^2 \hat{V}_T$$

$$\hat{V}_J = (1-f)\frac{n-1}{n}\sum_{j=1}^{n}(\hat{R}_{(j)} - \hat{R}_{(\cdot)})^2$$

où $\hat{R} = \bar{y}/\bar{x}$, $\hat{R}_{(j)} = \sum_{i\neq j} y_i / \sum_{i\neq j} x_i$ et $\hat{R}_{(\cdot)} = \sum_{j=1}^{n} \hat{R}_{(j)}/n$.

Soit $\hat{V}$ un de ces trois estimateurs de variance. Leblond (1989) démontre que, sous certaines conditions, le biais de $\hat{V}$ est alors

$$Biais(\hat{V}) = E\{\hat{V} - EQM(\bar{y}_Q)\}$$

$$= B_2 + O(n^{-3})$$

où $B_2 = O(n^{-2})$ est une expression complexe dépendant de y_k et où $O(\cdot)$ désigne que l'argument est borné. De plus Leblond (1989) démontre que pour $\hat{V}_T$, B_2 est toujours négatif et que pour $\hat{V}_J$, B_2 est toujours positif.

Nous proposons une méthode permettant de corriger $\hat{V}$ afin de réduire l'ordre du biais à un terme $O(n^{-3})$. Cette méthode consiste à obtenir un estimateur $\hat{B}_2$ de B_2 puis à définir l'estimateur de variance modifié $\hat{V}_M$ de la façon suivante:

(1) $$\hat{V}_M = \hat{V} - \hat{B}_2.$$

Toutefois, de par sa forme complexe, B_2 est difficile à estimer. Afin de simplifier l'estimation de B_2, nous considérons une approche assistée d'un modèle. Le modèle utilisé ici est celui généralement considéré comme sous-jacent à l'estimateur par le quotient:

$$E_\xi(y_i) = \beta x_i, \quad V_\xi(y_i) = \sigma^2 x_i \quad \text{et} \quad cov_\xi(y_i, y_j) = 0, \quad i \neq j.$$

Sous cette approche modéliste, nous considérons le biais anticipé de $\hat{V}$, soit

$$E_\xi[Biais(\hat{V})] = E_\xi(B_2) + O(n^{-3}).$$

L'avantage d'utiliser la notion de biais anticipé réside dans la forme simple de $E_\xi(B_2)$. En effet $E_\xi(B_2) = K_x\sigma^2$ où $K_x = O(n^{-2})$ ne dépend que des valeurs de x (voir tableau 1).

Tableau 1 **Biais asymptotique (jusqu'au terme $O(n^{-3})$) des estimateurs de variance.**

Biais	$\hat{V}_T$	$\hat{V}_{T2}$	$\hat{V}_J$	Estimation
B_2	<0	?	>0	complexe
$E_\xi(B_2)$	$\dfrac{-2M_x}{(1-f)}$	$\dfrac{-M_x(1+f)}{(1-f)}$	M_x	simple

où $M_x = \dfrac{(1-f)^2 S_x^2}{n(n-1)\overline{\overline{X}}}\sigma^2$.

Ainsi on obtient $\hat{V}_M$ en posant dans (1) $\hat{B}_2$ égal à l'estimateur de $E_\xi(B_2)$. Pour obtenir cet estimateur il suffit d'estimer σ^2. On exigera d'estimer σ^2 par $\tilde{\sigma}^2$ de telle sorte que

(2) $$\tilde{\sigma}^2 = C_x\hat{V}$$

et

(3) $$E_\xi(\tilde{\sigma}^2) = \sigma^2 + O_p(n^{-1})$$

où C_x est un terme ne dépendant que des valeurs de x et O_p désigne que l'argument est borné en probabilité par rapport au plan SI. La condition (2) permet d'obtenir un estimateur de variance modifié, $\hat{V}_M$, de forme simple puisqu'égal facteur multiplicatif × estimateur traditionnel. Afin de déterminer C_x pour chacun des trois estimateurs considérés, on peut démontrer le résultat suivant:

$$E_\xi\left[\left(\frac{\overline{\overline{X}}}{\overline{x}}\right)^q \hat{V}_T\right] = \sigma^2 C^{-1}_{x(q)} + O_p(n^{-2}).$$

où $C^{-1}_{x(q)} = \dfrac{(1-f)}{n-1}\left(\dfrac{\overline{\overline{X}}}{\overline{x}}\right)^q \overline{x}$

Pour $\hat{V} = \left(\frac{\bar{x}}{\bar{\bar{x}}}\right)^g \hat{V}_T$, on obtient alors $C_x = C_{x(g)}$ permettant de vérifier les équations (2) et (3):

$$C_{x(g)} = \frac{1}{C^{-1}_{x(g)}} = \frac{(n-1)}{\bar{\bar{x}}(1-f)}\left(\frac{\bar{x}}{\bar{\bar{X}}}\right)^g.$$

Le terme C_x est alors obtenu pour $\hat{V}_T$ et $\hat{V}_{T2}$ en posant respectivement $g = 0$ et $g = 2$ dans $C_{x(g)}$. La valeur de C_x pour $\hat{V}_J$ est la même que pour $\hat{V}_{T2}$ car

$$\hat{V}_{T2} = \hat{V}_J + O_p(n^{-2}).$$

Posant $\hat{B}_2 = C_x K_x \hat{V}$ dans (1), alors on obtient

$$\hat{V}_M = \hat{V} - K_x C_x \hat{V} = A_x \hat{V}.$$

où $A_x = 1 - K_x C_x$ est appelé facteur de correction. On a que A_x converge en probabilité vers 1. Ainsi asymptotiquement l'estimateur modifié converge vers sa version originale. La forme explicite des estimateurs de variance modifié des trois estimateurs ici considérés est alors obtenue en substituant les valeurs appropriées de K_x et C_x:

$$\hat{V}_{MT} = \left(1 + \frac{2}{n}(CV_x)^2 \frac{\bar{X}}{\bar{\bar{x}}}\right)\hat{V}_T$$

$$\hat{V}_{MT2} = \left(1 + \frac{1+f}{n}(CV_x)^2 \frac{\bar{x}}{\bar{\bar{X}}}\right)\hat{V}_{T2}$$

$$\hat{V}_{MJ} = \left(1 - \frac{1-f}{n}(CV_x)^2 \frac{\bar{x}}{\bar{\bar{X}}}\right)\hat{V}_J$$

où $CV_x = S_x/\bar{X}$ est le coefficient de variation de x dans la population. Remarquons que pour $\hat{V}_{MT}$ le facteur de correction est supérieur à 1 et que pour $\hat{V}_{MJ}$ ce facteur est inférieur à 1, ce qui devrait permettre, conformément aux résultats présentés dans le tableau 1, une réduction dans le bon sens du biais.

De plus, dans le cas où $x_i = 1$ pour tout i, les trois estimateurs traditionnels sont alors égaux à

$$\hat{V}(\bar{y}) = \frac{1-f}{n}\sum_{i=1}^{n}\frac{(y_i - \bar{y})^2}{n-1}.$$

Ainsi pour une telle situation l'estimateur de variance modifié est égal à l'estimateur sans biais habituellement utilisé. Ceci est un signe du bien fondé de la correction apportée.

Remarquons finalement que le biais B_2 d'ordre n^{-2} de $\hat{V}$ n'est pas complètement éliminé par la correction apportée. En effet cette correction s'effectue en soustrayant à $\hat{V}$ l'estimateur asymptotiquement sans biais de $E_\xi(B_2)$. Or pour une population quelconque $E_\xi(B_2)$ et B_2 ne sont pas nécessairement égaux (sauf si la population est totalement conforme au modèle ξ utilisé). Toutefois les résultats de simulation présentés dans la section suivante indiquent que dans la plupart des cas étudiés le facteur de correction permet une réduction substantielle du biais des estimateurs de variance traditionnels.

3 Résultats de simulations

Le but de l'étude par simulations était de comparer les estimateurs de variance modifiés et traditionnels présentés à la section 2. Le biais et l'erreur quadratique moyenne (EQM) sous certains modèles de superpopulation ont servi de critères de comparaison. Pour ce faire, nous avons considéré les quatre superpopulations correspondant au modèle suivant:

$$(4) \qquad E_\xi(y_i) = \beta x_i,\ V_\xi(y_i) = \sigma^2 x_i^\gamma \text{ et } cov_\xi(y_i, y_j) = 0,\ i \neq j$$

avec $\gamma = 0,\ 0.5,\ 1,\ 2$. Pour chacune de ces superpopulations, nous avons généré 60 populations finies de taille $N = 100$ à l'aide d'une fonction de densité gamma. L'utilisation d'une densité gamma est appropriée parce qu'elle est de distribution asymétrique tout comme la majorité des populations étudiées dans un contexte pratique. Pour chacune de ces 60 populations de chaque superpopulation nous avons généré 2,000 échantillons selon un tirage aléatoire simple sans remise de taille fixe, où pour chaque échantillon, l'estimateur par le quotient de même que divers estimateurs de variance étaient calculés. Nous avons considéré les tailles d'échantillon $n = 5, 10, 15, 20$.

Soient $\bar{y}_{Qrst}$ et $\hat{V}_{rst}$ respectivement l'estimateur par le quotient et un estimateur de variance calculés à partir du r^e échantillon de la s^e population de la t^e superpopulation. Soit également $\bar{Y}_{st}$ la moyenne de la s^e population de la t^e superpopulation.

L'EQM de l'estimateur par le quotient, le biais et l'EQM de $\hat{V}$ pour la s^e population de la t^e superpopulation sont respectivement définies par

$$EQM_{st}(\bar{y}_Q) = \sum_{r=1}^{2000} \frac{(\bar{y}_{Qrst} - \bar{Y}_{st})^2}{2000},$$

$$Biais_{st}(\hat{V}) = \sum_{r=1}^{2000} [\hat{V}_{rst} - EQM_{st}(\bar{y}_Q)]/2000.$$

$$Var_{si}(\hat{V}) = \sum_{r=1}^{2000}\left[\hat{V}_{rsi} - \left(\sum_{r=1}^{2000}\hat{V}_{rsi}/2000\right)\right]^2/1999.$$

Ensuite pour chaque superpopulation t on détermine le biais relatif et l'EQM de $\hat{V}$ par

$$Brel_i = \sum_{s=1}^{60}\frac{Biais_{si}(\hat{V})/EQM_{si}(\bar{y}_Q)}{60},$$

$$EQM_i(\hat{V}) = \sum_{s=1}^{60}\frac{(Biais_{si}(\hat{V}))^2 + VAR_{si}(\hat{V})}{60}.$$

Ces dernières quantités représentent une approximation de l'espérance sous le modèle de superpopulation i du biais relatif et de l'EQM des estimateurs de variance considérés. Elles servent de base de comparaison entre les différents estimateurs et sont présentées dans les tableaux 2 à 4.

Les populations générées à partir du modèle de superpopulation défini en (4) par $\gamma = 1$ correspondent à ce que nous appelons des populations favorables aux estimateurs de variance modifiés du fait que ces derniers sont obtenus à partir de ce même modèle. Ainsi pour de telles populations, on s'attend à ce que la correction apportée aux estimateurs de variance élimine le biais B_2 d'ordre n^{-2}. Par opposition, nous appelons populations non favorables les populations issues des trois autres modèles de superpopulation (avec $\gamma = 0, 0.5, 2$). Ces populations non favorables servent d'assises à la vérification de la robustesse des estimateurs de variance modifiés par rapport à la réduction du biais.

Les résultats présentés dans le tableau 2 illustrent que pour une population favorable les trois estimateurs de variance modifiés possèdent un biais d'ordre inférieur à leur version traditionnelle. Ceci se vérifie par le fait que pour une valeur de n fixé, le biais de $\hat{V}$ est supérieur en valeur absolue à celui de $\hat{V}_M$ et que le biais de $\hat{V}_M$ décroit plus rapidement que celui de $\hat{V}$. Ceci est vrai pour les trois estimateurs considérés. Remarquons également que contrairement à leur version traditionnelle, le biais des trois estimateurs modifiés est à peu près du même ordre de grandeur pour chaque valeur de n.

Tableau 2 Biais relatif des estimateurs de variance traditionnels et modifiés par rapport à la superpopulation favorable définie en (4) par $\gamma = 1$

n	$\hat{V}_T$	$\hat{V}_{MT}$	$\hat{V}_{T2}$	$\hat{V}_{MT2}$	$\hat{V}_J$	$\hat{V}_{MJ}$
5	-.326	-.065	-.183	-.046	.279	.086
10	-.181	-.022	-.108	-.021	.104	.016
15	-.118	-.003	-.069	-.003	.069	.013
20	-.092	-.004	-.056	-.003	.046	.006

Pour les populations non favorables, la correction permet une réduction significative du biais uniquement pour $\hat{V}_T$ et $\hat{V}_J$ (voir tableau 3). L'estimateur $\hat{V}_{T2}$ ne partage pas cette propriété de robustesse: pour plusieurs populations non favorables la correction a tendance à faire augmenter la valeur absolue du biais. Ceci peut s'expliquer par le fait que la forme même de cet estimateur est au départ liée au modèle de superpopulation favorable. En effet, $\hat{V}_{T2}$ est approximativement sans biais sous ce modèle.

Tableau 3 Biais relatif des estimateurs de variance traditionnels et modifiés par rapport à la superpopulation non favorable définie en (4) par $\gamma = 0$

n	$\hat{V}_T$	$\hat{V}_{MT}$	$\hat{V}_{T2}$	$\hat{V}_{MT2}$	$\hat{V}_J$	$\hat{V}_{MJ}$
5	-.378	-.112	-.057	.083	.421	.230
10	-.190	-.024	-.019	.073	.139	.053
15	-.119	-.000	-.007	.061	.082	.027
20	-.087	.004	-.005	.050	.056	.018

Au tableau 4, nous présentons les résultats concernant l'EQM des estimateurs de variance pour une des superpopulations considérés. On y remarque que les estimateurs de variance modifiés $\hat{V}_{MT}$ et $\hat{V}_{MT2}$ possèdent, règle générale, une EQM supérieure à leur version traditionnelle alors que $\hat{V}_{MJ}$ possède une EQM inférieure à celle de $\hat{V}_J$. Si l'on compare entre eux l'EQM des estimateurs de variance, alors on obtient

$$EQM(\hat{V}_T) < EQM(\hat{V}_{T2}) < EQM(\hat{V}_J)$$

avec les mêmes relations pour les modifications de ces estimateurs. Ces constatations s'appliquent également aux autres superpopulations considérés lors des simulations.

Tableau 4 EQM des estimateurs de variance traditionnels et modifiés par rapport à la superpopulation définie en (4) par $\gamma = 0.5$

n	$\hat{V}_T$	$\hat{V}_{MT}$	$\hat{V}_{T2}$	$\hat{V}_{MT2}$	$\hat{V}_J$	$\hat{V}_{MJ}$
5	36.8	62.1	311.4	357.2	1134.1	968.6
10	2.9	3.8	9.0	10.1	12.1	10.8
15	0.6	0.8	1.4	1.6	1.7	1.6
20	0.21	0.25	.42	.45	.46	.47

En résumé, comme estimateur de variance fiable et robuste (c'est à dire avec biais relativement peu élevé, EQM faible et ce, le plus indépendamment possible de la distribution de la population considérée) nous recommandons, à la lumière des résultats obtenus, l'utilisation de $\hat{V}_{MT}$. D'autre part, si on désire utiliser un estimateur jackknife, alors on a tout à gagner à utiliser $\hat{V}_{MJ}$, car celui-ci possède à la fois un biais et une EQM inférieur à $\hat{V}_J$.

4 Conclusion

Nous avons présenté une méthode visant à réduire le biais des estimateurs de variance traditionnellement utilisés pour estimer la variance de l'estimateur par le quotient. Cette méthode consiste à multiplier ces estimateurs de variance par un facteur de correction permettant ainsi une réduction du biais. Les résultats de simulation démontrent que cette correction peut s'avérer efficace et robuste lorsqu'appliquée aux estimateurs de variance de Taylor et jackknife. Pour ce dernier estimateur, la correction a pour effet de diminuer son EQM.

Il serait intéressant de généraliser la méthode présentée ici à l'estimation de variance d'autres types d'estimateur de régression: combiné, quotient séparé, régression généralisé, etc..

Bibliographie

Leblond, Y. (1989), *Contributions à la théorie d'estimation pour sous-populations*, Thèse de doctorat, Université de Montréal.

Connaissez-vous S ?

Normand Ranger
Laboratoire de statistique-informatique
Département d'informatique et de recherche opérationnelle
Université de Montréal

1. Avant-propos

Le statisticien d'aujourd'hui utilise abondamment les logiciels statistiques dans le cadre de ses travaux. Quelques logiciels réputés ($SPSS^X, SAS, BMDP$, etc.) sont fort connus. Il nous apparaît important de présenter un autre logiciel de plus en plus populaire: S. Celui-ci présente des particularités intéressantes qui en font une alternative à considérer parmi l'ensemble des outils statistiques.

2. Introduction

2.1 Définitions

Certains nommeraient S un logiciel, progiciel. D'autres diraient qu'il est un langage de programmation. Les concepteurs le présentent comme *un environnement de programmation pour l'analyse de données et le graphisme*... Quoiqu'il en soit, ce *logiciel* (appelons-le ainsi) possède certaines caractéristiques particulières:

- interactif dans un environnement UNIX;
- langage fonctionnel:
  - tout calcul est effectué par des fonctions;
  - le résultat d'un calcul est un élément (complexe...);
- utilise le concept de l'objet.

2.2 Concepts

Toute manipulation se fait sur un objet qui est défini par un ensemble de caractéristiques propres à celui-ci:

nom

type : vecteur, matrice, tableau, catégorie, série chronologique, liste, fonction, etc.;

caractéristiques typiques : nom(s), longueur (vecteur), dimensions (matrice, tableau), période (série), etc.

2.3 Liens hors logiciel

Symbiose possible avec d'autres éléments du système informatique.

- système d'exploitation UNIX:
  - commandes UNIX exécutables durant une session S;
  - utilisation d'éditeurs (ex.: ed, vi) pour éditer les objets S;
- programmes FORTRAN et C exécutables à travers les fonctions S;
- utilisation de S en programmation C;
- utilisation de librairies (ex.: IMSL).

3. Objets S

3.1 Manipulations

Plusieurs types de manipulations sont possibles.

lire un objet, lister son contenu;

transformer à l'aide d'opérations arithmétiques;

extraire et créer ainsi d'autres objets:
- vecteur extrait d'une matrice de données;
- sous-matrice extraite d'une matrice de données;

calculer à l'aide de fonctions (procédures) qui produiront des résultats.
(Ex.: régression, histogramme, etc.)

Note: tout objet transformé (créé, modifié) durant une session S est conservé sur disque dans un répertoire particulier toujours accessible durant toute session S (présente et future).

3.2 Matrice

Très souvent les données à analyser peuvent être représentées par une matrice. Exemple: [1]

```
 900  525 3571 1976  541
 900  572 4092 1250  524
 900  580 3865 1586  561
 750  529 4870 2351  414
 800  544 4399  431  410
1000  571 5342 1333  457
 800  451 5319 11868 344
 800  553 5126 2138  467
 800  529 4447 8577  464
 700  552 4512 8507  498
```

3.2.1 Exemples de manipulations

- créer l'objet S (matrice) à partir du fichier de données sur disque:
  $essence < -matrix(scan("/usr/ranger/ess.d"), ncol = 5, byrow = T)$
- associer les caractéristiques à cet objet:
  $dimnames(essence) < -list(nom.etat, c("tax", "dlic", "inc", "road", "fuel"))$
- consulter l'objet:
  - au complet [2]
    essence

```
    tax dlic  inc  road fuel
ME  900  525 3571  1976  541
NH  900  572 4092  1250  524
VT  900  580 3865  1586  561
MA  750  529 4870  2351  414
RI  800  544 4399   431  410
CN 1000  571 5342  1333  457
NY  800  451 5319 11868  344
NJ  800  553 5126  2138  467
PA  800  529 4447  8577  464
OH  700  552 4512  8507  498
```

[1] Nous présentons seulement les 10 premières données d'un ensemble de données réelles tiré de: Weisberg, S., Applied Linear Regression, 2nd edition, John Wiley & Sons, 1985, p. 34-36.

[2] Voir remarque 1.

– une donnée en particulier:
essence["*MA*", *c*("*tax*", "*fuel*")]

```
tax fuel
750  414
```

– et de façon équivalente mais moins poétique...
essence[4, *c*(1, 5)]

- transformer, créer un autre objet:
ess.log < −*log*(*essence*)
ess.log[3]

```
          tax     dlic      inc     road     fuel
ME 6.802395 6.263398 8.180601 7.588830 6.293419
NH 6.802395 6.349139 8.316789 7.130899 6.261492
VT 6.802395 6.363028 8.259717 7.368970 6.329721
MA 6.620073 6.270988 8.490849 7.762596 6.025866
RI 6.684612 6.298949 8.389133 6.066108 6.016157
CN 6.907755 6.347389 8.583355 7.195187 6.124683
NY 6.684612 6.111467 8.579041 9.381601 5.840642
NJ 6.684612 6.315358 8.542081 7.667626 6.146329
PA 6.684612 6.270988 8.399985 9.056839 6.139885
OH 6.551080 6.313548 8.414496 9.048645 6.210600
```

Note: on remarque que le nouvel objet créé (*ess.log*) a les mêmes caractéristiques (type, noms, dimensions, etc.) que l'objet de départ (*essence*).

- créer un autre objet (vecteur) à partir de la matrice:
taxe.ess < −*essence*[, "*tax*"]
taxe.ess

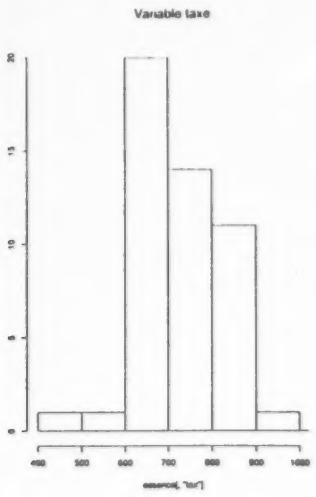
```
 ME  NH  VT  MA  RI  CN  NY  NJ  PA  OH  IN  IL  MI  WI  MN  IA  MO  ND  SD
900 900 900 750 800 1000 800 800 800 700 800 750 700 700 700 700 700 700 700

 NE  KS  DE  MD  VA  WV  NC  SC  GA  FA  KY  TN  AL  MS  AR  LA  OK  TX  MT
850 700 800 900 900 850 900 800 750 800 900 700 700 800 750 800 658 500 700

 ID  WY  CO  NM  AZ  UT  NV  WN  OR  CA
850 700 700 700 700 700 600 900 700 700
```

- effectuer des calculs complexes:

– sur un sous-ensemble (colonne, variable) de l'objet:

[3] Voir remarque 1.

$hist(essence[, \text{“}tax\text{”}], main = \text{“}Variable\ taxe\text{”})$



- sur l'objet *manipulé* sous un autre angle:
  $hist(essence)$[4]
- sur une fonction d'un objet:
  $hist(abs(log(essence[, \text{“}tax\text{”}])))$

3.3 Fonction et liste

3.3.1 Fonction

La **fonction** qui est un des objets les plus fondamentaux, est équivalente dans d'autres langages ou logiciels à une **macro**, une **procédure**, une **sous-routine**, un **sous-programme** ou à une **fonction**...

L'usager peut:

- utiliser les fonctions existantes (corrélations, graphiques, analyse discriminante, etc.);
- créer ses propres fonctions (en utilisant les fonctions existantes) et y associer une documentation interactive;

[4]Une des particularités de *S* est de pouvoir manipuler un objet sous divers aspects. Dans cet exemple, la fonction *hist* s'attend à avoir comme premier argument un vecteur. L'objet *essence* étant une matrice, la fonction produira l'histogramme de toutes les colonnes de la matrice juxtaposées une à la suite de l'autre (donc un vecteur). Dans certaines situations, cette manipulation peut être intéressante...

- modifier facilement toute fonction existante dans le logiciel.

Le résultat d'une fonction élaborée (ex. régression) sera:

- immédiatement accessible à l'écran (entre autres pour les graphiques);
- à conserver dans un objet qui sera très souvent une liste.

3.3.2 Liste

La liste est un objet qui est constitué d'objets de tout type: constante, vecteur, matrice, liste (qui contiendra d'autres objets!), etc.

3.3.3 Exemple

Soit *resul.reg* l'objet résultant de l'appel d'une fonction de régression:

$$resul.reg < -lsfit(x, y, yname = "ACFAS")$$

L'objet *resul.reg* sera une liste qui contiendra entre autres:

- un vecteur *coef* contenant les paramètres estimés;
- un vecteur *residuals* contenant les résidus;
- une clé logique *intercept* indiquant la présence de l'ordonnée à l'origine;
- une liste *qr* contenant d'autres vecteurs, matrices, valeurs (dont le rang), etc.;
- un titre.

Cet objet *resul.reg* sera utilisé dans d'autres fonctions pour compléter l'analyse.
Exemples:
$plot(resul.reg\$residuals, y, main = "Residus\ et\ y")$
$print(resul.reg\$coef)$
$qqnorm(resul.reg\$residuals)$

3.4 Autres objets

Mentionnons qu'il existe d'autres objets de grande utilité dont:

tableau à plusieurs dimensions (pour ANOVA)

série l'objet de type série chronologique contient les caractéristiques propres à lui et certaines fonctions sont propres à ce type d'objets.
Exemples:
$serieA < -ts(scan("/home/ranger/sA"), start = c(1960, 1), freq = 12)$
$window(serieA, start = c(1970, 3), end = c(1975, 3))$
$tsplot(serieA, main = "ACFAS(bis)")$

opérateurs arithmétiques Le logiciel S considère un opérateur arithmétique comme une fonction. On peut donc créer de nouveaux opérateurs. De plus, les opérateurs arithmétiques tiennent compte du type des objets. Ainsi, il existe un produit, un produit vectoriel, produit matriciel, etc.

etc.

3.5 Librairie

Il existe une possibilité de créer des librairies de nouvelles fonctions sur un sujet particulier. Ces librairies peuvent être écrites par l'usager ou celui-ci peut avoit accès à une multitude de librairies publiques disponibles dans des groupes d'usagers (S-News, Statlib). Ces librairies sont accessibles par courrier électronique.

4. Graphiques

Un des aspects les plus intéressants de S tient à ses capacités graphiques. Mentionnons seulement que:

- les fonctions graphiques produisent, par défaut, de "beaux" graphiques;
- plusieurs fonctions graphiques permettent "d'ajouter" interactivement à un graphique existant:
  - titres, légendes;
  - autres graphiques superposés;
- on peut identifier, avec la souris, des points d'un graphique;
- il y a des outils simples pour redéfinir une fenêtre (feuille) graphique:
  - marges, échelles, axes, taille des caractères, etc.;
  - nombre de graphiques par fenêtre (feuille).

5. Analyses possibles

Tout comme les autres logiciels, il existe une multitude de fonctions permettant de faire des calculs en statistique ou, de façon plus générale, en mathématiques et aussi de programmer comme dans un langage de programmation conventionnel. Mentionnons seulement quelques possibilités:

- **statistique (calcul)**
  - descriptif (moyenne, médiane, etc.);

- estimation de fonction de densité;
- analyse de séries chronologiques;
- fonctions de probabilité connues:
  - simulation (générer des valeurs aléatoires), densité, cumulative, quantile;
- ANOVA (*S* n'offre pas beaucoup dans ce domaine...);
- régression:
  - moindres carrés, meilleurs sous-ensembles, norme L1, robuste;
- analyse multivariée:
  - classification, corrélation canonique, analyse discriminante, modèle log-linéaire, arbre de recouvrement minimal, analyse en composantes principales, etc.

• **statistique (graphique)**

- histogramme, diagramme en secteurs, nuage de points;
- diagramme en boîtes;
- lissage de courbes;
- quantiles normalisés (*qqnorm plot*);
- faces de Chernoff, étoiles et autres représentations graphiques multivariées.

• **mathématiques**

- algèbre matricielle: produit croisé, produit matriciel, valeurs et vecteurs propres, etc.;
- résolution de système d'équations linéaires;
- approximation spline.

• **programmation**

- outils (instructions) classiques de programmation: IF, FOR, etc.;
- manipulation de fichiers;
- récursivité.

6. Avantages et désavantages

Rien n'est parfait! Mentionnons quelques aspects positifs et négatifs de ce logiciel.

Avantages

• orienté sur une approche interactive;

- accessibilité totale
  - résultats (tout peut être conservé);
  - fonctions (toute fonction du logiciel peut être modifiée);
- possibilité de conserver les résultats et autres objets créés;
- possibilité de programmation avancée et d'intégration de procédures écrites en d'autres langages
- lien étroit avec UNIX.

Désavantages

- peut exiger de la programmation (en *S*);
- environnement UNIX;
- absence de fonctions dans des domaines fondamentaux (ex.: ANOVA);
- présentation faible des résultats;
- lourdeur de manipulation pour les valeurs manquantes.

7. Considérations techniques

Il serait trop long de décrire tous les aspects techniques du logiciel, mentionnons seulement:

- coût $\approx$ 1 200$ US;
- espace-disque nécessaire $\approx$ une trentaine de Mo;
- installation:
  - environnement UNIX;
  - peut être implanté sur plusieurs modèles d'ordinateurs;
  - peut profiter des capacités offertes par les programmes de gestion d'écran (SUNTOOLS, X).

De plus, dans cet article nous avons présenté la version New*S* du logiciel. Il existe aussi une version plus ancienne (old*S*) et une version *augmentée* (*S*+).
Pour plus de détails, il serait bon de consulter le manuel de référence [5] ou de communiquer

[5] Becker, R.A., Chambers, J.M. et Wilks, A.R., The New S Language, Wadsworth & Brooks/Cole, Pacific Grove, California, 1988.

avec les concepteurs [6] .

8. Conclusion

Sans être le logiciel statistique idéal (en existe-t-il un?), *S* peut être envisagé comme:

- très bon outil graphique;
- très utile pour la recherche;
- bon complément à d'autres logiciels statistiques.

[6] Adresse et téléphone:
AT&T Software Sales
P.O. Box 25000
Greensboro, North Carolina 27420
(800) 828-UNIX

NOUVELLE ASSOCIATION DE STATISTICIENS

Christian Desbiens, statisticien
Service de la statistique et de l'économie des transports
Ministère des Transports du Québec

Présenté dans le cadre du 58e congrès de l'ACFAS
Université Laval, mai 1990

Note: Le générique masculin est employé dans le but d'alléger le texte.

Nouvelle association de statisticiens

1. Introduction

Tout corps d'emploi qui se respecte se doit d'être regroupé en société, corporation ou autre. Mais il est difficile de satisfaire tous les professionnels d'un même corps d'emploi à l'aide d'une seule association. De plus, lorsque cette profession a peu de chance de pouvoir former une corporation, il devrait exister une association qui puisse réagir aux problèmes auxquels les professionnels sont confrontés. Elle doit être mandatée pour faire respecter la profession et défendre les intérêts des professionnels. Pour cette raison et quelques autres, je vais tenter de vous sensibiliser sur la nécessité pour les statisticiens de former cette nouvelle association. Il sera aussi question de ce que seront ses objectifs, ses avantages et conséquences et finalement quelles seront son organisation et sa composition.

2. Pourquoi une nouvelle association de statisticiens

Toute tentative de se regrouper en corporation aurait pour fruit une désillusion complète. Le chemin pour y parvenir contient un abîme profond. Les conditions d'évaluation pour former un tel regroupement sont exigeantes. Il semble que les critères requis ne sont pas remplis par notre profession. Etre reconnu comme oeuvrant dans un domaine exclusif n'étant pas possible par un regroupement en corporation, une association pallierait à ce manque.

Un autre besoin impérieux qui serait satisfait si on regroupait les statisticiens, est la promotion de la statistique comme une science avant-gardiste et le statisticien comme étant le seul spécialiste de cette science. A tout le moins, si ce n'est pas le statisticien qui agit à titre de spécialiste, il faudrait que la personne oeuvrant avec cette science ait prouvé qu'elle possède les connaissances suffisantes. Pour cette raison, une association regroupant tous les spécialistes de cette science serait un atout pour parvenir à réaliser le besoin.

Le statisticien dans notre société n'est pas très visible. Ce problème d'image est dû à l'absence d'une organisation pouvant canaliser toutes les énergies pour y remédier. L'association proposée aura donc à créer la crédibilité professionnelle du statisticien ainsi que de promouvoir les possibilités de la statistique. Lors du dernier concours de recrutement de statisticiens pour oeuvrer dans la Fonction publique du Québec, les exigences d'éligibilité au concours étaient très larges. Ce concours permettait à presque tout le monde de postuler et, éventuellement, d'agir à titre de statisticien. Si l'association avait déjà été mise sur pied, elle aurait pu protester. Elle aurait fait valoir les droits des professionnels de la statistique en réclamant des exigences plus strictes pour participer à ce concours (i.e. un baccalauréat en statistique ou mathématiques, un baccalauréat avec une maîtrise en statistique, etc.). Ceci n'est qu'un des nombreux exemples de la nécessité d'une telle

association. Une inertie de notre part ne fera que faire durer cette image effacée que nous projetons dans le public en général.

L'information qui circule en statistique est souvent en anglais, théorique et traite principalement de sujets pointus s'éloignant du même coup de la réalité de tous les jours. Il y a trop de statisticiens qui travaillent en statistique appliquée, qui rencontrent fréquemment des problèmes d'ordre pratique pour qu'une association les néglige. De l'information pratique, en français et sur des sujets courants, serait bienvenue. Une association proche de ses membres et de leurs besoins pourrait combler cette lacune.

Il n'y a pas d'association locale ou d'activité d'échange rapprochant les statisticiens oeuvrant dans un même milieu leur permettant de profiter de l'expérience des collègues. Ainsi, la création d'une association pourrait aisément, encore une fois, régler ce problème.

Donc, une telle association, priorisant ces points, réglerait bien des problèmes. Mais, puisqu'elle n'existe pas encore, il faudra la créer et la former en travaillant sur la science elle-même, le professionnel et finalement le public en général.

3. Principaux objectifs visés

Chaque association a un nombre d'objectifs et de mandats dont seulement une faible proportion est réalisée. La réalisation de ceux-ci montre la priorité qu'une association établit au niveau de ses mandats. L'association proposée ici déterminera certains buts à atteindre, n'ayant pas tous la même importance. Voici donc les principaux objectifs que cette association pourraient se fixer:

- diffuser, commenter et valoriser l'information. Cette information pourrait être de nouvelles méthodes d'analyse, des logiciels de pointe, des nouvelles sur les activités du monde de la statistique...;
- créer une bibliographie la plus complète possible en statistique, à laquelle tous les membres auraient accès;
- faciliter les échanges entre les statisticiens.

Cependant, les objectifs qui suivent sont également intégrés à la charte de chaque association sans être nécessairement mis en pratique.

- mettre l'accent sur le français;
- promouvoir le métier de statisticien;
- contribuer à l'avancement de la statistique;

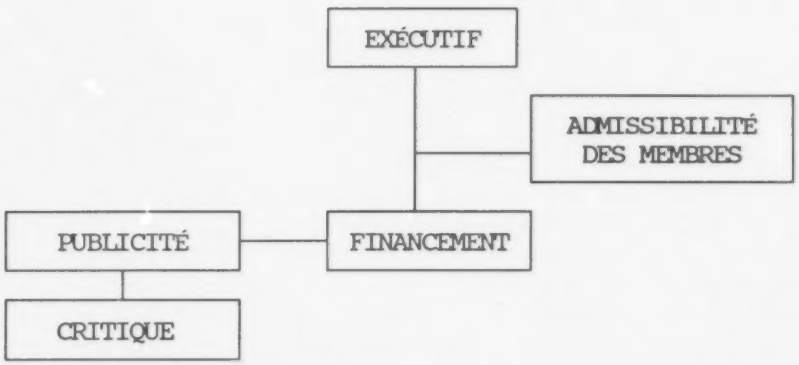
- défendre la statistique et la profession;
- instruire le public des capacités et des limites de la statistique;
- améliorer la qualité de l'information diffusée au public;
- critiquer l'information diffusée de façon erronée (revue, média...).

4. Composition et Organisation

Compte tenu des objectifs que l'association veut se donner, il pourrait y avoir des critères de sélection des membres. Evidemment, de tels critères ne seraient définitifs que lorsque le Conseil exécutif serait en place. Voici, par exemple, comment on pourrait aborder le problème. Il serait peut-être favorable de n'accepter comme membre, que des personnes ayant prouvé leur compétence. En ce sens, un examen d'entrée pour les gens n'ayant pas de formation en statistique serait requis. Cet examen pourrait être fait par un groupe de personnes désignées par l'exécutif. Dans un autre scénario, il pourrait s'avérer que les gens ayant une formation en statistique ou mathématiques n'aient pas à passer cet examen. Toute hypothèse n'est, pour le moment, que spéculation, le point de l'éligibilité demeurant litigieux et demandant une concertation.

En plus de combler des vides importants dans plusieurs domaines, l'association offrirait une gamme de services sur mesure pour ses membres. Parmi les privilèges, il pourrait y avoir: une carte de membre, une revue, une journée de congrès, quelques séminaires...

La structure organisationnelle de l'association pourrait avoir l'allure suivante:



Le conseil exécutif verrait à gérer tous les mandats de l'association. Le conseil pour l'admissibilité des membres serait temporaire de sorte qu'il serait créé selon le besoin. Le conseil de la publicité aurait la charge de

sensibiliser le public sur la statistique à tout point de vue. La critique serait étroitement liée à la publicité car elle aurait le mandat de freiner les abus. Comme la publicité, elle aidera à nous rendre plus visible. Finalement, le financement s'occupera de gérer les coffres de l'association. Dans ce cas aussi, tout serait à décider par l'assemblée des membres lors de la création de cette association.

5. Conclusion

Dans le cas où la démonstration de l'importance de la création d'une association ayant les grandes orientations précédentes vous a convaincu, je vous invite à vous joindre à moi pour la créer. Sans l'effort de tous et l'adhésion d'une majorité de statisticiens, cette association ne restera qu'une idée, attrayante peut-être, mais qu'une idée.

L'UNION FAIT LA FORCE!

Mario Montégiani*

RÉSUMÉ

Au Québec, depuis quelques années, il est possible d'observer une augmentation du nombre de personnes exerçant la profession de statisticien. Cette augmentation a eu des effets bénéfiques sur la qualité et la quantité d'analyses et d'informations statistiques produites. Toutefois, elle n'a malheureusement pas eu les mêmes effets sur la reconnaissance de la profession et du professionnel de la statistique. Bien qu'aujourd'hui, pour la majorité des professions, les professionnels d'un même corps d'emploi adhèrent à un regroupement, les statisticiens demeurent encore isolés et bien peu renseignés sur les associations susceptibles de les sortir de cette solitude.

L'adhésion massive de leur part à un de ces groupements permettra à chacun d'entre eux, et à la profession, de trouver la place qui leur est due.

(Note: Il est à noter que l'emploi du générique masculin est utilisé ici dans le but d'alléger le texte.)

---

* Mario Montégiani est statisticien au Service de la statistique et de l'économie des transports au Ministère des Transports du Québec

L'UNION FAIT LA FORCE!

La session à laquelle vous assistez aujourd'hui a pour but de vous faire prendre conscience de l'importance des regroupements professionnels. Quatre personnes convaincues de cette importance expliqueront leur vision d'un regroupement des professionnels de la statistique. Ainsi, cette session se veut une tribune où les personnes qui ont des choses à dire pour regrouper le plus de statisticiens dans des associations de professionnels ont une chance de s'exprimer. Pour leur part, elles sont toutes membres d'au moins une association et elles vont parler d'options possibles pour ceux qui voudraient devenir membres d'une ou de plusieurs associations.

Déroulement de la session

L'ordre dans lequel ils présenteront leur conférence a été déterminé au hasard par un tirage qui s'est déroulé juste avant le début de la session. Ainsi l'ordre établi sera:

* M. Victor Tremblay;

* Mad. Pascale Rousseau;

* pause

* M. Christian Desbiens;

* M. Christian Genest.

Comme dans toutes les autres sessions, une période de question d'une durée maximale de 5 minutes suivra chaque communication. Lorsque toutes les personnes auront présenté leur communication, ils viendront tous à l'avant, et, dans le même ordre que la présentation des conférences, auront deux minutes chacun pour résumer leurs idées. Suivra un débat où tous pourront participer en posant des questions ou en faisant des commentaires sur ce qui vous intéresse plus particulièrement.

Ce forum permettra à tous les participants de questionner ou de commenter les différentes possibilités suggérées par les conférenciers. Chacun aura ainsi la chance d'exprimer son propre point de vue sur l'ensemble du sujet. Pour ma part, j'exposerai succinctement les avantages d'un regroupement de professionnels.

Qu'est-ce qu'une association?

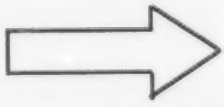
Au Québec, on retrouve plus de 31 000(1) associations qui ont des buts très différents. Dans la majorité des cas, ces associations ont pour premier objectif de regrouper des personnes ayant des activités, des idées ou des points en commun. L'idée de ce genre de regroupement existe depuis plusieurs siècles et a été éprouvée plusieurs fois. Peu importe la source de motivation des premières personnes à s'être regroupées en association, les résultats escomptés étaient : une économie de temps et d'argent , la chance de tirer profit de la somme des expériences de cette collectivité et d'utiliser l'importance du groupe pour que les diverses positions adoptées par cette association en profite.

Quelque soit le créneau de l'association, les bénéfices retombent toujours sur les membres. Son adhésion en fait une personne qui a: une influence directe sur les positions adoptées par son association, un groupe de pression pouvant promouvoir et à l'occasion défendre les causes qui lui sont chères et une assemblée partageant une même perspective.

Aussi, pour toutes ces raisons, cette session se veut le choc initiateur d'un mouvement secouant le milieu de la statistique pour qu'une très grande majorité de statisticiens se réunissent au sein d'une association correspondant à leurs intérêts et à ceux de la statistique.

Pourquoi une association de professionnels de la statistique?

TOUT



SOMME DES PARTIES

«La somme des parties est inférieure au tout...»
Gestalt allemande

Figure 1

---

1 : Source Inspecteur général des institutions financières du Québec

Quelle que soit l'importance et la valeur de toutes les personnes qui ont un point en commun, l'attention que leur portera l'ensemble de la population sera certainement moins importante que si elles sont réunies pour faire un groupe homogène, une association.

Ainsi, le plus brillant des médecins aurait sûrement moins d'écoute de la part de la population s'il est seul à prendre position sur un sujet quelconque que si la corporation des médecins se prononce.

Ce comportement est d'autant plus observable quand la profession et la science n'ont pas acquis leurs lettres de noblesse. Ce qui est encore le lot du statisticien et de la statistique. Car, malgré l'emploi sans cesse croissant de l'un et l'autre, encore bien peu de personnes en reconnaissent l'importance. Sans dire ou penser qu'une association pourrait régler tous les problèmes de reconnaissance qu'éprouvent tous ceux qui pratiquent la statistique, elle pourrait, à tout le moins, en régler un certain nombre.

En fait, l'idée que les gens d'une même profession se regroupent pour former une association n'est pas nouvelle. Au Québec, il existe même un code des professions obligeant certains professionnels à être membres d'une corporation pour avoir le droit d'exercer ou de porter le titre rattaché à leur spécialisation.

Cette obligation pour certains professionnels de se regrouper sous la tutelle d'une association officialisée par la loi, vient de la volonté de protéger le grand public face à la possible incompétence de toute personne se prétendant spécialiste. Cette même loi oblige les corporations à évoluer dans un cadre très rigide. Ainsi plusieurs professionnels doivent activement s'impliquer dans l'administration et les différents comités obligatoires à la formation d'une telle corporation. Elles réservent à leurs membres le travail spécialisé pour lequel ils ont reçu une formation académique. Elles sont aussi les principales organisations prenant parole dans les débats concernant leurs membres et leurs travaux. Par exemple, quand il est question du système judiciaire ou de toute nouveauté dans ce domaine le Barreau est un interlocuteur privilégié.

Mais, malgré tout le charme que certaines personnes trouvent à ces corporations, le fardeau de l'ensemble des obligations est lourd à porter. Même que certaines corporations se plaignent de la rigidité du cadre dans lequel elles évoluent. Et c'est sans doute pour ces raisons, malgré certains avantages évidents, que l'Office des professions du Québec a refusé les 27 dernières demandes de constitution de corporation qui lui ont été présentées. Plutôt que de s'enfermer dans ce carcan, elle suggère de créer une association regroupant tous (autant que faire se peut) les professionnels oeuvrant dans le domaine.

Donc, il semble clair que la voie à suivre réside entre la situation actuelle et une corporation. Il y a donc une foule de possibilités avantageuses pour la statistique et les statisticiens. Car bien que la science de la statistique ait connu quelques gains dans les dernières années, il reste encore beaucoup de place pour l'amélioration. Ainsi, si tous

les statisticiens devenaient membres d'une association, il s'établirait un réseau de connaissances permettant une meilleure circulation de l'information entre les professionnels. De ceci, découlerait une meilleure efficacité et, par conséquent, plus de valeur accordée à la statistique et au statisticien.

Allons encore plus loin et supposons que tous les statisticiens deviennent membre d'une seule association. Cette supposition a pour effet de consacrer cette association comme le représentant officiel, devant tous les intervenants majeurs d'une société (gouvernement, groupe de pression, public, ...), des statisticiens et de leurs intérêts. Quel force pour un groupe de professionnels d'avoir une voix parlant avec l'appui de tous les professionnels du domaine. De plus cette supposition implique que cette association aurait des moyens d'une ampleur considérable permettant ainsi à tous ses membres de voir leurs besoins professionnels comblés d'une façon adéquate, ce qui n'est pas possible si l'association n'a que 10 membres. Finalement, cette hypothèse constituerait un réseau complet de ressources pouvant améliorer l'ensemble des problèmes de la société.

Mais qu'on soit rêveur ou réaliste, un seul fait demeure. Tout professionnel, quel qu'il soit et quel que soit son domaine d'application, aurait tout avantage à faire partie d'une organisation lui permettant de rester à la fine pointe de son domaine. Son appartenance lui assurerait une certaine participation sur les actions qu'elle met en oeuvre pour promouvoir les bienfaits et limites de sa science.

Situation dans le monde de la statistique

Il existe en statistique des associations qui ont pour but de regrouper tous les professionnels du domaine. Le problème est qu'encore bien des statisticiens manquent à l'appel. La plus importante au Canada, avec 774 membres en 1989, est sans doute la Société statistique du Canada (SSC). La majorité de ses membres viennent d'institutions académiques, ce qui en fait avant tout une société savante de statistique. Cette association possède quelques associations régionales qui relèvent directement de la SSC. La Société statistique de Montréal est une de ces associations régionales.

Il existe aussi à travers le monde, des associations d'envergure nationale. Presque tous les pays ont au moins une association nationale qui permet aux statisticiens de ces pays d'être regroupés, comme il se doit. Nos voisins du sud, les américains, n'y font pas exception. Ils ont l'équivalent de la SSC avec l'American Statistical Association (ASA). Plus vieille et plus grande, l'ASA a des moyens peu commun pour une association de statisticiens. Elle édite plusieurs revues et compte un nombre très impressionnant de plus de 15 000 membres. Plusieurs de ces membres viennent de l'étranger, faisant de cette association la plus internationale des associations nationales.

D'autres associations ont un caractère international, on n'a qu'à penser à: l'Institut international de statistique (IIS), l'Association Internationale des statisticiens d'enquête (AISE), International Association of Statistical Computing (IASC), etc. ... Ces associations couvrent le monde entier et ont pour but de rassembler les statisticiens de tous les pays de façon à ce que le monde statistique évolue ensemble et qu'il n'y ait pas de cloison inutile. Différents buts sont poursuivis par ces associations, mais le plus important est toujours celui de faire circuler l'information sur les toutes dernières nouveautés de la statistique.

Pour la majorité des associations précédemment nommées, leurs membres n'ont d'autre obligation que de payer leur cotisation pour en devenir ou en rester membres. Par contre, pour certaines autres associations, la seule façon d'en devenir membre est d'être élu lors d'une élection où seul les membres déjà élus peuvent voter (e.g. l'IIS). Ce dernier genre d'association assure donc une qualité et une compétence exceptionnelle des membres. Évidemment ,chacun des types d'accès a ses avantages et ses inconvénients, mais chose assurée il existe actuellement des associations pour presque tous les goûts, il suffit tout simplement d'y adhérer.

Aperçu de la tendance à l'association chez des statisticiens

Mais pour certains professionnels qui participent à ce congrès, quelle est la situation? Le questionnaire (c.f. annexe 1) qui vous était proposé tout au long de la semaine se voulait une façon de connaître cette situation pour les personnes ayant suffisamment de curiosité pour prendre le questionnaire. D'aucune façon les réponses à ce questionnaire ne se veulent les résultats d'un sondage ayant comme population les participants au colloque sur les Méthodes et domaines d'application de la statistique. Il ne se veut que le recensement de l'opinion d'une quarantaine (42) de personnes curieuses ayant le goût de s'exprimer sur le sujet (les données servant à cette analyse se trouve à l'annnexe 2). Parmi les personnes ayant répondu à ce questionnaire 45% personnes étaient membres d'associations. Parmi les non-membres, une proportion importante de ceux-ci (88%) seraient prêts à devenir membres d'une association s'ils pouvaient en trouver une qui corresponde vraiment à leurs besoins.

Pour ce qui est des personnes déjà membres d'association qui ont répondu au questionnaire, le tableau suivant montre les proportions selon les divers degrés de satisfaction.

	Totalement satisfait				Totalement insatisfait
	1	2	3	4	5
Répartition des répondants	7%	33%	47%	13%	0%

Tableau 1 - Degré de satisfaction des membres d'association

Évidemment, puisque personne n'est obligé de demeurer membre à vie d'une association, il est normal que personne ne soit totalement insatisfait. Par contre, ce qui est surprenant, c'est qu'il semble, malgré des cotisations assez élevées, ne pas y avoir une totale satisfaction envers leurs associations.

Donc, d'un côté les membres répondants ne sont pas totalement satisfait. D'un autre côté, plusieurs non-membres répondants, malgré leur intention de venir membre d'une association, ne le sont pas. Comment expliquer cette situation? C'est peut-être que les associations ne s'occupent pas des objectifs recherchés?

Dans le questionnaire sept (7) objectifs étaient présentés et il était demandé aux répondants de les ordonner selon leurs préférences (1 étant ce qu'ils préfèrent le plus et 7 étant ce qu'ils préfèrent le moins).

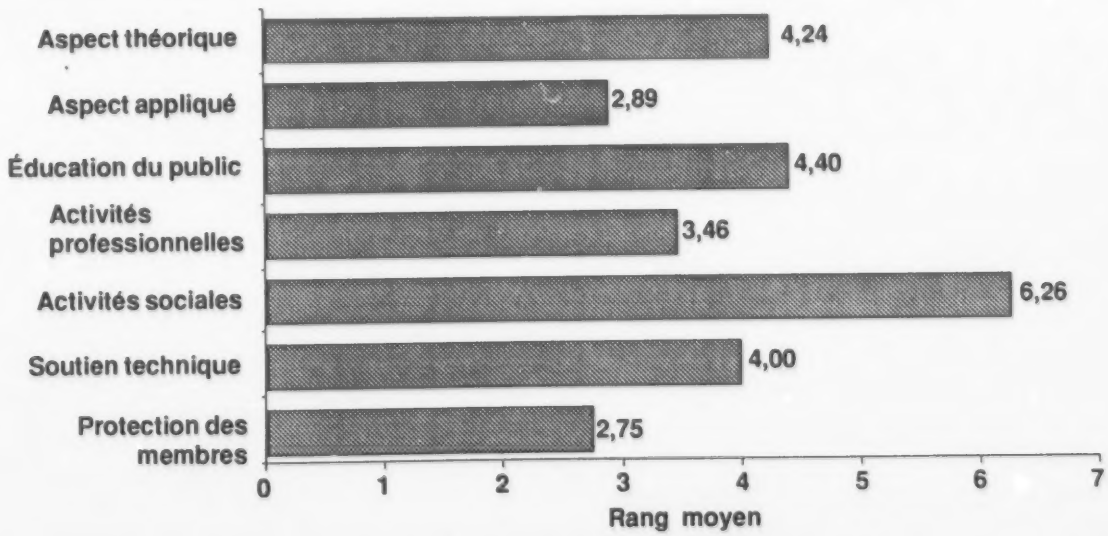


Figure 2 - Classements des objectifs

Ainsi, si les activités sociales n'ont pas la faveur populaire, une association qui aurait comme objectifs de:

- Protéger ses membres et l'intégrité de la profession en dénonçant les utilisations abusives de la statistique;

- Renseigner ses membres sur les récents développements de la statistique appliquée;

serait très populaire auprès de ceux qui ont répondu au questionnaire. Il est à noter aussi que malgré un désintéressement marqué pour les activités sociales, les répondants considèrent que les activités professionnels (colloque, séminaire, congrès,...) sont très importantes. Les objectifs concernant le soutien technique à fournir à ses membres, les récents développements théoriques de la statistique et l'éducation du public sur l'expertise statistique apparaissent un peu plus secondaire pour ces mêmes personnes.

Pour ce qui est du point de vue des répondants sur l'accessibilité à une association de statisticiens, les avis sont assez partagés.

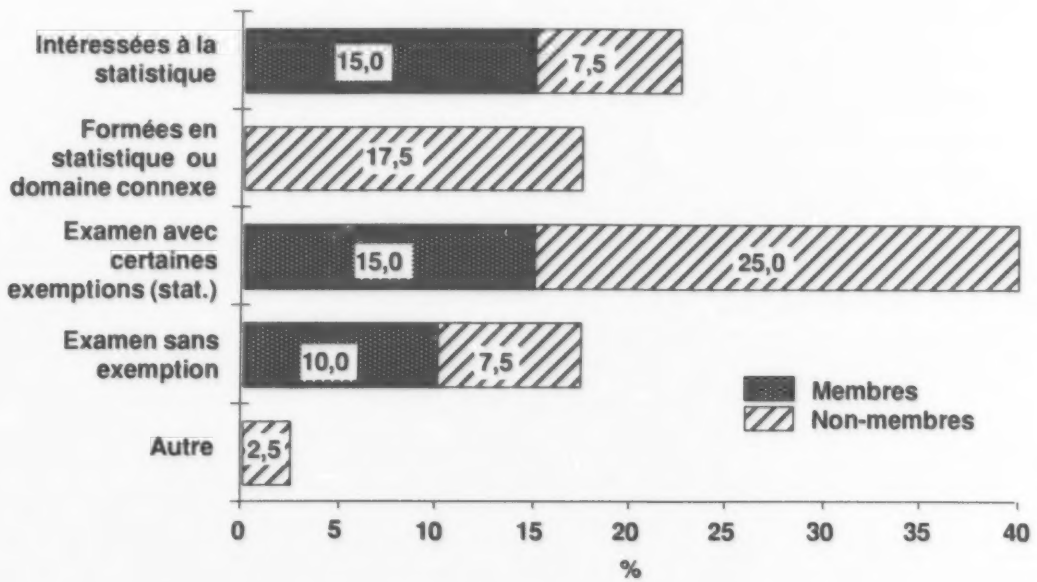


Figure 3 - Accessibilité à une association de statisticiens

Puisque la majorité des associations ont comme politique d'accepter toutes les personnes payant leur cotisation, il est possible qu'un bon nombre de répondants éprouvent une certaine insatisfaction vis-à-vis l'accessibilité à une association professionnelle.

Conclusions

Même si l'utilisation des méthodes statistiques et l'emploi du statisticien ne cessent d'augmenter, toute entreprise pour augmenter leur importance et leur valeur doit être mise en oeuvre. Une des plus efficaces et des plus facilement réalisable consiste à ce que tous les statisticiens se regroupent à l'intérieur d'associations. Si chacun des professionnels de la statistique fait partie d'au moins une de ces organisations, il ne sera plus isolé et son adhésion permettra ainsi d'accroître la visibilité de son association et, par conséquent, de la statistique.

Dans une étape subséquente, il pourrait être intéressant de réunir tous ces professionnels dans une seule association pour ainsi créer un consensus efficace pour la promotion de toutes activités reliées à l'avancement de la statistique.

Évidemment, nous ne sommes pas encore rendus à cette étape mais si on veut y arriver, il est important que chacun apporte sa contribution dès maintenant, car pour pouvoir courir il faut tout d'abord savoir marcher.

Questionnaire sur les associations professionnelles de statistique

Espace réservé à la codification

Identification :

Nombre d'années d'expérience de travail en statistique

Secteur de travail

Diplôme(s) acquis (cochez la ou les cases appropriées)

				statistique	mathématiques	autres domaines
moins d'un an	☐ 1	gouv. provincial	☐ 1	certificat ☐ 01	☐ 02	☐ 03
entre 1 et 3	☐ 2	gouv. fédéral	☐ 2	bacc. ☐ 04	☐ 05	☐ 06
entre 3 et 5	☐ 3	enseig. universitaire	☐ 3	maîtrise ☐ 07	☐ 08	☐ 09
entre 5 et 10	☐ 4	enseig. collégial	☐ 4	doctorat ☐ 10	☐ 11	☐ 12
10 et plus	☐ 5	secteur privé	☐ 5			
		étudiant(e)	☐ 6			
		autre	☐ 7			

Question 1.
Voici sept propositions sur les objectifs possibles d'une association de professionnel(le)s de la statistique. Ordonnez-les selon vos priorités. Si plusieurs éléments sont d'égale importance, mettez-leur un rang identique.

_____ Renseigner ses membres des récents développements théoriques de la statistique

_____ Renseigner ses membres des récents développements de la statistique appliquée

_____ Éduquer le public en l'informant et en vulgarisant les diverses techniques d'expertise statistique

_____ Favoriser les échanges entre les membres par diverses activités professionnelles (colloque, séminaire, congrès,...)

_____ Favoriser les échanges entre les membres par diverses activités sociales

_____ Offrir un soutien technique aux membres en leur permettant de profiter de certains services améliorant leur travail (évaluation de livres, évaluation de logiciels, etc. ...)

_____ Protéger ses membres et l'intégrité de la profession en dénonçant les utilisations abusives de la statistique.

Question 2.
Choisissez la proposition correspondant le mieux à l'idée que vous avez de l'accessibilité à une association de statisticien(ne)s.

☐ 1 Être accessible à toute personne intéressée à la statistique.

☐ 2 Être accessible uniquement aux personnes ayant eu une formation en statistique (détentrices d'un diplôme) ou dans un domaine connexe (mathémathique, économie, épidémiologie,...)

☐ 3 Être accessible à toute personne qui réussit un examen évaluant ses compétences avec exemption pour celle ayant reçu une formation en statistique (détentrice d'un diplôme).

☐ 4 Être accessible à toute personne qui réussit un examen évaluant ses compétences

☐ 5 Aucune des propositions précédentes. Mon idée correspond plutôt à: _____

_____

_____

Partie pour les personnes qui sont déjà membres d'une ou de plusieurs associations de statisticien(ne)s

Question 3.

a. Indiquez si vous êtes membre des associations suivantes en cochant la case correspondante au nom de l'association

1. Société Statistique du Canada (SSC) ☐
2. American Statistical Association (ASA) ☐
3. Association internationale des statisticiens d'enquête (AISE) ☐
4. Institut international de statistique (IIS) ☐
5. Société statistique de Montréal (SSM) ☐
6. International Association of Statistical Computing (IASC) ☐
7. Biometric Society ☐
8. Autre: _____ ☐

b. Sur une échelle de 1 à 5 donnez votre satisfaction globale envers la ou les associations dont vous êtes membre.

satisfaction 1 2 3 4 5 insatisfaction

c. Combien allouez-vous, par année, en droit d'inscription?

☐1 moins de $ 30 ☐2 entre $ 30 et $ 60
☐3 entre $ 60 et $ 90 ☐4 entre $ 90 et $120
☐5 entre $120 et $150 ☐6 plus de $150

d. Contribuez-vous aux activités d'une des associations dont vous êtes membre, en vous engageant dans l'organisation de celle-ci?

☐1 oui ☐2 non

Partie pour les personnes qui ne sont pas membres d'une association de statisticien(ne)s

Question 3.

Seriez-vous intéressé(e) a devenir membre d'une association de professionnel(le)s en statistique?

☐1 oui ☐2 non ——>le questionnaire se termine ici

Question 4.

a. Combien seriez-vous prêt à allouer, par année, en droit d'inscription?

☐1 moins de $ 30 ☐2 entre $ 30 et $ 60
☐3 entre $ 60 et $ 90 ☐4 entre $ 90 et $120
☐5 entre $120 et $150 ☐6 plus de $150

b. Seriez-vous intéressé(e) à consacrer de votre temps pour contribuer aux activités de votre association en vous engageant dans l'organisation de celle-ci?

☐1 oui ☐2 non ☐3 ne sait pas

Espace réservé à la codification

17 23

24

25

26

27

28

29

Commentaires:

_____

_____

_____

_____

_____

ANNEXE 2

Jeu de données recueilli lors du colloque sur les méthodes et domaines d'application de la statistique 1990

AN_EXP	TRAV	SCOL	THEORIE	APPLIQU	PUBLIC	CONGRES	SOCIAL	SERVICE	PROTEGER	ACCES	MEMB	SATISF	PRIX_M	ACT_M	FUTUR_M	PRIX_NM	ACT_NM
1	1	000004	5.0	3.0	4.0	6.0	7.0	2.0	1.0	2	.	.	.	.	1	3	3
2	1	000004	5.0	1.5	6.0	4.0	7.0	3.0	1.5	3	.	.	.	.	1	3	3
2	1	000004	6.0	2.0	4.5	2.0	7.0	4.5	2.0	3	10000000	4	2	2	.	.	.
5	1	000509	2.0	1.0	4.0	5.0	7.0	3.0	6.0	4	10000000	3	4	2	.	.	.
4	1	000004	6.0	4.0	1.0	2.0	3.0	6.0	6.0	1	10000000	3	2	2	.	.	.
5	.	000006	2.0	1.0	4.0	3.0	7.0	5.0	6.0	1	00001000	3	1	2	.	.	.
3	2	000507	4.0	5.0	2.0	6.0	7.0	3.0	1.0	3	01100000	3	4	2	.	.	.
2	1	000004	2.0	3.0	4.0	6.0	7.0	5.0	1.0	2	.	.	.	.	1	4	1
2	1	000004	7.0	4.0	3.0	5.0	6.0	2.0	1.0	4	.	.	.	.	1	2	1
2	1	000004	3.5	3.5	2.0	5.0	7.0	6.0	1.0	3	.	.	.	.	1	4	3
2	5	000407	2.5	2.5	4.5	4.5	7.0	6.0	1.0	3	.	.	.	.	1	4	1
2	1	000004	5.0	5.0	2.0	5.0	7.0	3.0	1.0	3	.	.	.	.	1	4	1
3	1	000004	3.5	3.5	5.0	2.0	7.0	6.0	1.0	3	11000000	2	6	2	.	.	.
5	3	040710	5.0	5.0	5.0	1.0	2.0	5.0	5.0	1	10000011	.	.	1	.	.	.
5	1	000507	4.5	4.5	3.0	2.0	7.0	6.0	1.0	3	.	.	.	.	1	3	1
4	.	000005	2.5	2.5	7.0	1.0	6.0	4.0	5.0	3	.	.	.	.	1	1	3
2	2	000005	3.0	2.0	6.0	4.0	7.0	5.0	1.0	3	.	.	.	.	1	4	2
1	6	.	3.5	3.5	7.0	3.5	3.5	6.0	1.0	4	.	.	.	.	1	3	1
3	1	000004	3.0	3.0	6.0	3.0	7.0	5.0	1.0	3	10100110	4	6	2	.	.	.
3	1	000004	6.0	4.0	5.0	3.0	7.0	2.0	1.0	2	.	.	.	.	1	3	1
3	1	000004	4.0	4.0	4.0	1.5	6.0	1.5	7.0	5	.	.	.	.	2	.	.
4	2	000004	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.
3	2	000405	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.
3	1	000105	3.5	3.5	6.0	2.0	7.0	1.0	5.0	1	.	.	.	.	1	1	2
2	2	000004	5.0	1.0	2.5	2.5	4.0	7.0	6.0	3	10000000	3	3	2	.	.	.
2	2	000004	7.0	5.0	1.0	3.0	6.0	4.0	2.0	2	.	.	.	.	1	3	1
3	1	000607	6.0	1.0	5.0	2.0	7.0	3.0	4.0	2	.	.	.	.	1	3	3
2	1	000004	5.0	4.0	3.0	7.0	6.0	2.0	1.0	2	.	.	.	.	1	5	1
5	3	000507	6.5	3.0	5.0	1.0	6.5	2.0	4.0	2	.	.	.	.	1	6	3
4	1	000509	6.5	6.5	4.5	1.5	1.5	4.5	3.0	1	.	.	.	.	1	2	2
1	6	000004	4.0	2.0	7.0	5.0	6.0	3.0	1.0	4	10000000	3	1	2	.	.	.
3	5	000407	4.0	3.0	1.0	6.0	7.0	5.0	2.0	1	10000000	3	4	2	.	.	.
2	1	000004	1.5	1.5	5.0	6.0	7.0	4.0	3.0	3	.	.	.	.	1	1	3
4	1	000406	5.0	2.0	6.0	1.0	7.0	4.0	3.0	4	.	.	.	.	1	2	3
2	7	000407	1.5	1.5	6.0	5.0	7.0	3.5	3.5	1	10000000	2	2	2	.	.	.
2	1	000407	1.5	1.5	6.0	3.0	7.0	4.0	5.0	1	10000000	2	2	2	.	.	.
2	7	000407	6.0	1.5	3.5	5.0	7.0	1.5	3.5	4	10000000	2	1	2	.	.	.
2	2	000004	2.0	1.0	7.0	3.0	6.0	4.0	5.0	4	10000000	2	3	2	.	.	.
2	1	000407	3.0	4.0	7.0	2.0	6.0	5.0	1.0	3	10000000	1	1	2	.	.	.
2	1	000004	6.0	2.0	4.5	2.0	7.0	2.0	4.5	1	.	.	.	.	1	3	1
3	1	000004	5.5	2.0	3.5	3.5	7.0	5.5	1.0	3	.	.	.	.	1	3	2
3	1	000004	5.0	2.0	3.5	3.5	7.0	6.0	1.0	3	.	.	.	.	1	3	2

IASS International Association of Survey Statisticians
AISE Association Internationale des Statisticiens d'Enquêtes

Institut international
de statistique

Présentation de l'Association internationale des statisticiens d'enquêtes

Par : Victor Tremblay
Président de STATPLUS

Extrait du feuillet publicitaire de l'Association

L'Association internationale des statisticiens d'enquêtes

L'Association internationale des statisticiens d'enquêtes (I.A.S.S.-A.I.S.E.) a été créée en 1973 lors d'un congrès de l'Institut international de statistique dont elle est une des sections. Elle a pour objectif de promouvoir l'étude et le développement de la théorie et de la pratique des recensements et enquêtes statistiques. Elle cherche aussi à accroître l'intérêt pour ces sujets parmi les statisticiens, diverses organisations et institutions, les gouvernements et le public en général dans les différents pays du monde.

L'A.I.S.E. poursuit ces buts grâce aux moyens suivants :

- des réunions internationales qui se tiennent tous les deux ans conjointement avec celles de l'Institut international de statistique et qui sont un forum de discussion sur la statistique d'enquête. Elles comportent des sessions sur les développements récents dans les méthodes d'enquêtes et de recensements et sur leurs applications;

- des réunions régionales ou des séminaires consacrés à des sujets spécifiques;

- des publications envoyées systématiquement ou sur demande, et des facilités offertes pour l'abonnement à certaines revues à des tarifs réduits.

Services offerts par l'AISE

I. L'A.I.S.E. envoie gratuitement à tous ses membres :

- le «Statisticien d'enquêtes», qui est le journal de l'Association et qui paraît deux fois par an, en français et en anglais. On y trouve des informations générales sur la vie de l'Association (congrès, séminaires, ...) et sur les autres congrès et réunions à travers le monde pouvant intéresser les membres. Il comprend aussi diverses rubriques techniques sur des questions liées aux enquêtes : terminologie, bibliographie, résumés d'articles ou d'ouvrages;

- les «Informations internationales de statistique», qui est le bulletin d'information de l'Institut international de statistique et de ses cinq sections et qui est publié trois fois par an;

- un recueil des contributions présentées aux congrès de l'Institut international de statistique et se rapportant particulièrement aux enquêtes et recensements;

- un annuaire des membres de l'Association.

II. Les membres de l'Association peuvent aussi souscrire à des tarifs préférentiels des abonnements aux revues suivantes :

- la «Revue internationale de statistique» (revue de l'Institut international de statistique) qui traite des sujets pouvant intéresser les statisticiens de toutes disciplines;

- «Techniques d'enquêtes/Survey Methodology», publié par Statistique Canada qui donne deux fois par an des articles consacrés à la théorie et à la pratique des enquêtes;

- le «Journal of Official Statistics» (publié par Statistics Sweden) qui, dans quatre numéros annuels, traite de la méthodologie des opérations menées par les services officiels de statistique;

- «A Celebration of Statistics», le volume spécial édité en 1985 à l'occasion du centenaire de l'Institut international de statistique au prix de 26 $ US (au lieu de 39 $).

III. Les membres de l'A.I.S.E. souhaitant bénéficier de l'avis technique d'experts de l'Association en matière d'enquêtes ou de recensements ont la possibilité d'obtenir une réponse à leurs questions dans la rubrique «Question/Réponse» du «Statisticien d'enquêtes».

Adhésions

Les personnes désirant adhérer à l'A.I.S.E. sont invitées à s'adresser à :

Mme A.M. VESPA-LEYDER
Secrétaire de l'A.I.S.E.
c/o I.N.S.E.E.
18, bd Adolphe Pinard
75675 PARIS CEDEX 14 - FRANCE